Module 15: Multilevel Modelling of Repeated Measures Data

MLwiN Practical[1]

George Leckie, Tim Morris & Fiona Steele
Centre for Multilevel Modelling

Pre-requisites

- MLwiN practicals for Modules 3 and 5

If you find this module helpful and wish to cite it in your research, please use the following citation:

Leckie, G., Morris, T., Steele, F. (2016). Multilevel Modelling of Repeated Measures Data - MLwiN Practical. LEMMA VLE Module 15, 1-76. URL: http://www.bristol.ac.uk/cmm/learning/online-course/.

Contents

[1] This MLwiN practical is adapted from the corresponding Stata practical: Steele, F. (2014). Multilevel Modelling of Repeated Measures Data: Stata Practical. LEMMA VLE Module 15, 1-61. (http://www.bristol.ac.uk/cmm/learning/course.html).

P15.1 Repeated Measures Data

P15.1.1 Introduction to physical health functioning dataset

In the first part of this practical we will fit growth curve models to data on health functioning from a study of British civil servants called the Whitehall II study (also known as the Stress & Health Study).[2] Health functioning was assessed by the SF-36, a 36 item instrument that comprises eight subscales covering physical, psychological and social functioning. These eight scales can be summarised into physical and mental health components. These are scaled using general US population norms to have mean values of 50 and low scores imply poor functioning. We will study change in physical health functioning which was measured on up to six occasions for each respondent.

The data are in wide form, i.e. with one record per individual and six variables for health functioning at the six measurement occasions. The dataset also includes information on the respondent's age at each occasion, their employment grade at the first occasion, and their gender. The analysis file contains the following variables for 4427 individuals:

Variable	Description and codes
id	Individual identifier (coded 1, 2, . . ., 8815)
female	Gender (1=female, 0=male)
grade	Employment grade at baseline (1=high, 2=intermediate, 3=low)
age1	Age at occasion 1 (years)
phf1	Physical health functioning score at occasion 1
…	…
age6	Age at occasion 6 (years)
phf6	Physical health functioning score at occasion 6

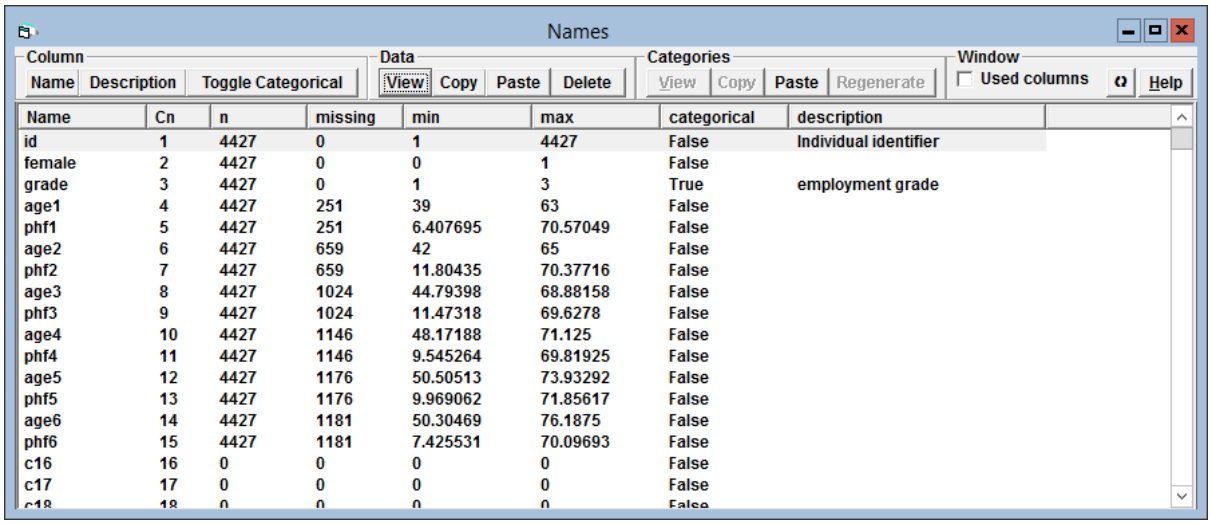
[2] http://www.ucl.ac.uk/whitehallII

To open the worksheet:

From within the LEMMA Learning Environment

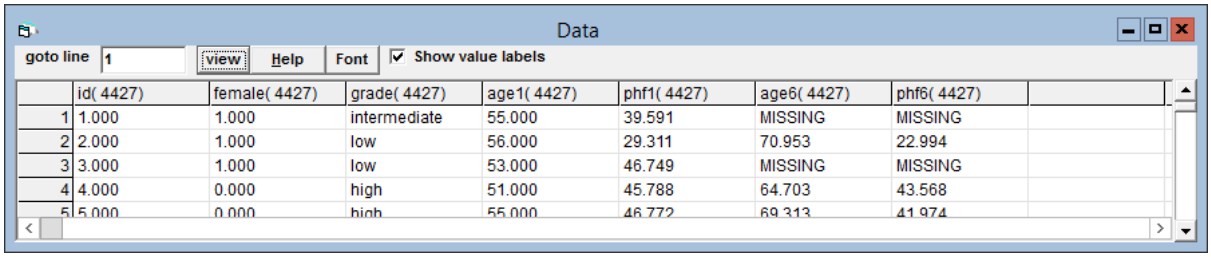
- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "15.1.wsz"

The **Names** window will appear.



Name	Cn	n	missing	min	max	categorical	description
id	1	4427	0	1	4427	False	Individual identifier
female	2	4427	0	0	1	False	
grade	3	4427	0	1	3	True	employment grade
age1	4	4427	251	39	63	False	
phf1	5	4427	251	6.407695	70.57049	False	
age2	6	4427	659	42	65	False	
phf2	7	4427	659	11.80435	70.37716	False	
age3	8	4427	1024	44.79398	68.88158	False	
phf3	9	4427	1024	11.47318	69.6278	False	
age4	10	4427	1146	48.17188	71.125	False	
phf4	11	4427	1146	9.545264	69.81925	False	
age5	12	4427	1176	50.50513	73.93292	False	
phf5	13	4427	1176	9.969062	71.85617	False	
age6	14	4427	1181	50.30469	76.1875	False	
phf6	15	4427	1181	7.425531	70.09693	False	
c16	16	0	0	0	0	False	
c17	17	0	0	0	0	False	

- To view a selection of the data, in the **Names** window select the variables **id**, **female**, **grade**, **age1**, **phf1**, **age6**, and **phf6** and click the **View** button under the **Data** heading.



	id(4427)	female(4427)	grade(4427)	age1(4427)	phf1(4427)	age6(4427)	phf6(4427)
1	1.000	1.000	intermediate	55.000	39.591	MISSING	MISSING
2	2.000	1.000	low	56.000	29.311	70.953	22.994
3	3.000	1.000	low	53.000	46.749	MISSING	MISSING
4	4.000	0.000	high	51.000	45.788	64.703	43.568
5	5.000	0.000	high	55.000	46.772	69.313	41.974

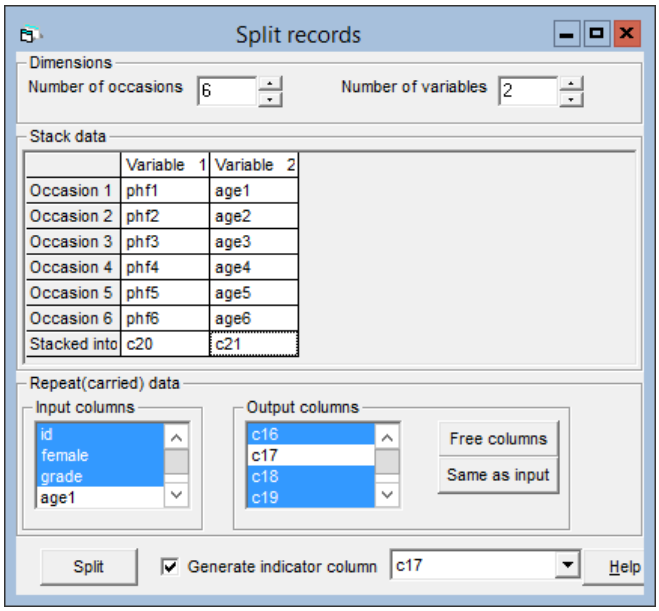
We can see that individuals 1 and 3 have missing data for occasion 6. We will obtain a summary of missing data patterns before fitting any models.

P15.1.2 Restructuring data from wide to long form

Our first task is to convert the data from wide form (i.e., one record per individual) to long form (i.e., one record per occasion per individual). In other words, we need to create six records for each individual, with each record corresponding to one of the six measurement occasions of the two variables **age** and **phf**.

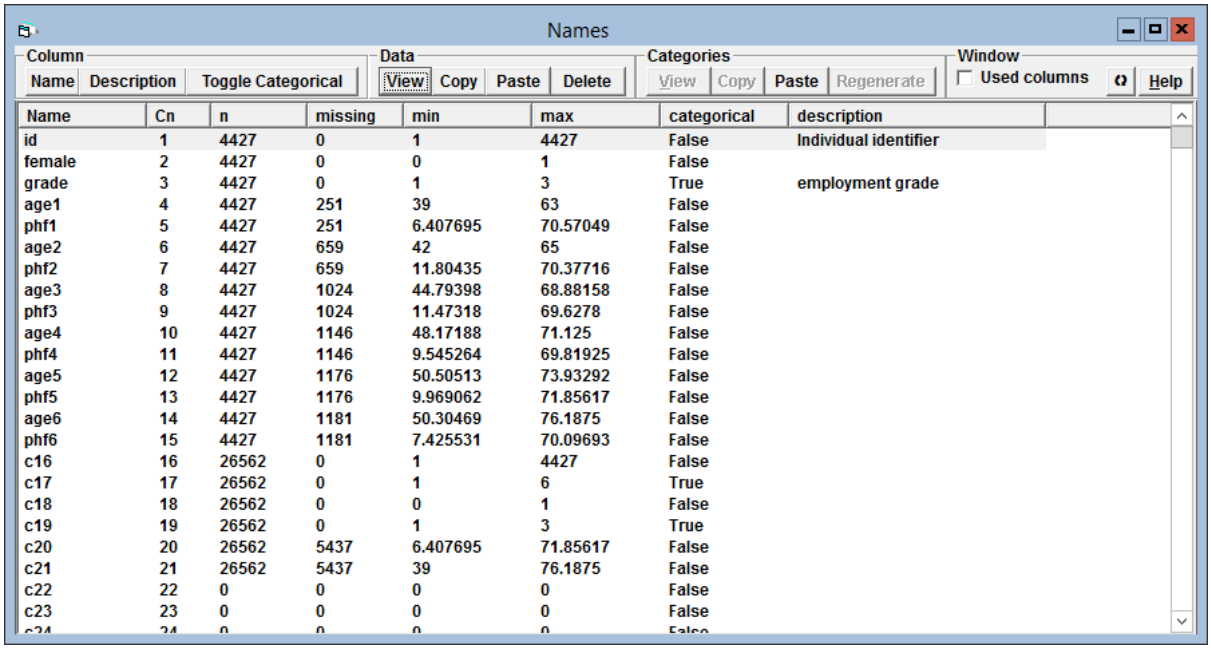
- From the **Data Manipulation** menu, select **Split Records**
- Change **Number of occasions** to **6**
- Change **Number of variables** to **2**
- Under **Stack data**, use the mouse pointer to fill the first column with the variables **phf1** through to **phf6** and then **c20**, and the second column with the variables **age1** through to **age6** and then **c21**. This tells the window to generate a new variable containing the occasion specific values of **phf** in column **c20** and the corresponding occasion specific values of **age** in column **c21**. These two columns therefore have one record per occasion per individual.
- Under **Repeat(carried) data**, select **id**, **female** and **grade** as the **Input columns**, and **c16**, **c18** and **c19** as the **Output columns**. This tells the window to create long-form versions of the variables **id**, **female**, and **grade** in columns **c16**, **c18**, and **c19**, respectively.
- Finally, check the **Generate indicator column** checkbox and select **c17** in the drop down list to create a measurement occasion indicator variable in **c17**.

Your **Split records** box should look as follows:



- Click **Split**, and when prompted to save the dataset select **No**

Our dataset now stores the original data in both wide and long form. This is confirmed when we inspect the **Names** window. The window shows the original wide-form variables are of length 4427 while the new long-form versions of these variables are of length 26562 (= 6 × 4427).



Name	Cn	n	missing	min	max	categorical	description
id	1	4427	0	1	4427	False	Individual identifier
female	2	4427	0	0	1	False	
grade	3	4427	0	1	3	True	employment grade
age1	4	4427	251	39	63	False	
phf1	5	4427	251	6.407695	70.57049	False	
age2	6	4427	659	42	65	False	
phf2	7	4427	659	11.80435	70.37716	False	
age3	8	4427	1024	44.79398	68.88158	False	
phf3	9	4427	1024	11.47318	69.6278	False	
age4	10	4427	1146	48.17188	71.125	False	
phf4	11	4427	1146	9.545264	69.81925	False	
age5	12	4427	1176	50.50513	73.93292	False	
phf5	13	4427	1176	9.969062	71.85617	False	
age6	14	4427	1181	50.30469	76.1875	False	
phf6	15	4427	1181	7.425531	70.09693	False	
c16	16	26562	0	1	4427	False	
c17	17	26562	0	1	6	True	
c18	18	26562	0	0	1	False	
c19	19	26562	0	1	3	True	
c20	20	26562	5437	6.407695	71.85617	False	
c21	21	26562	5437	39	76.1875	False	
c22	22	0	0	0	0	False	
c23	23	0	0	0	0	False	

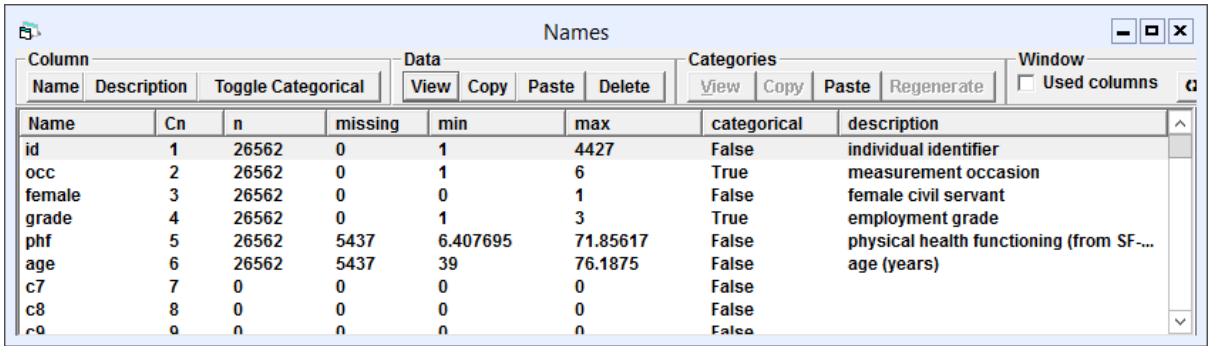
In the interests of keeping the worksheet as simple as possible, we will remove all wide-form variables as we will not analyse these further.

- From the **Data Manipulation** menu, select **Command interface**
- In the text entry area at the bottom of the **Command interface** window, type the following commands one by one, pressing return after each one has been inputted to run the commands:

```
ERAS C1-C15
MOVE
NAME C1 'id'
NAME C2 'occ'
NAME C3 'female'
NAME C4 'grade'
NAME C5 'phf'
NAME C6 'age'
DESC 'id' 'individual identifier'
DESC 'occ' 'measurement occasion'
DESC 'female' 'female civil servant'
DESC 'grade' 'employment grade'
DESC 'phf' 'physical health functioning (from SF-36)'
DESC 'age' 'age (years)'
```

- These commands will remove (**ERAS**) the wide-form variables in columns **c1** to **c15**, shift (**MOVE**) all remaining variables (i.e., variables currently in columns **c16** to **c21**) to the now empty columns at the beginning of the worksheet, rename (**NAME**) the six long-form variables, and lastly add variable labels (**DESC**) to these six variables

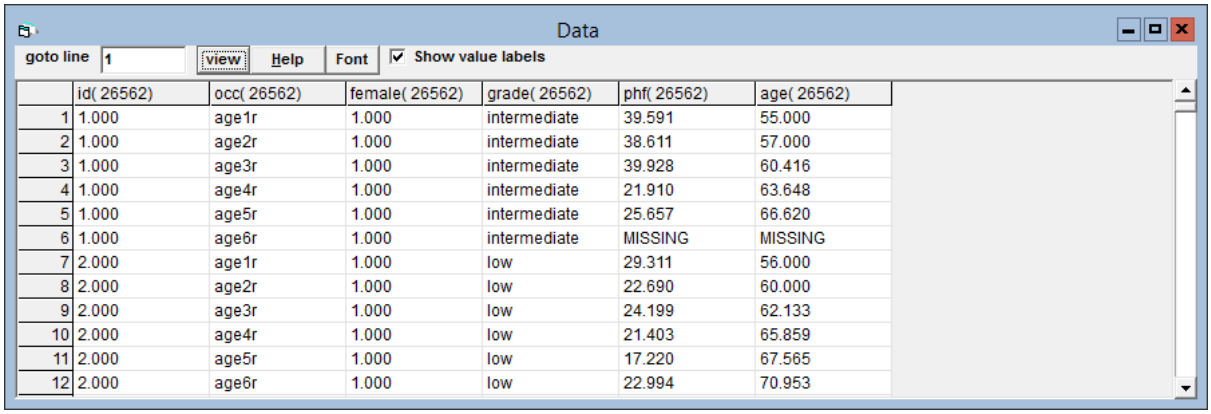
Your **Names** windows should now look as follows:



Name	Cn	n	missing	min	max	categorical	description
id	1	26562	0	1	4427	False	individual identifier
occ	2	26562	0	1	6	True	measurement occasion
female	3	26562	0	0	1	False	female civil servant
grade	4	26562	0	1	3	True	employment grade
phf	5	26562	5437	6.407695	71.85617	False	physical health functioning (from SF-...
age	6	26562	5437	39	76.1875	False	age (years)
c7	7	0	0	0	0	False	
c8	8	0	0	0	0	False	
c9	9	0	0	0	0	False	

Note that the restructured file now only contains the six long form variables **id**, **occ**, **female**, **grade**, **age** and **phf**. Each variable has 26562 records which is six times the number of individuals in the data.

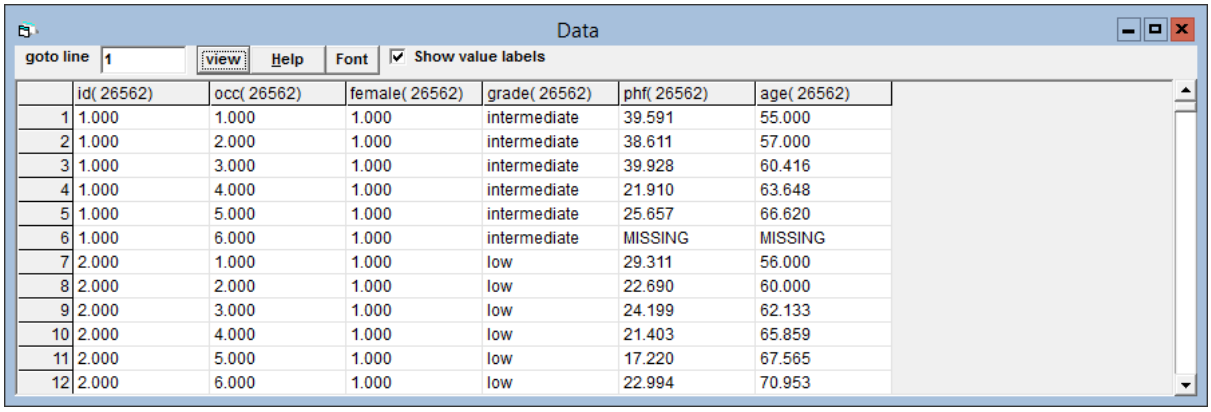
- Select these six variables in the **Names** window and click the **View** button to view the new dataset



	id(26562)	occ(26562)	female(26562)	grade(26562)	phf(26562)	age(26562)
1	1.000	age1r	1.000	intermediate	39.591	55.000
2	1.000	age2r	1.000	intermediate	38.611	57.000
3	1.000	age3r	1.000	intermediate	39.928	60.416
4	1.000	age4r	1.000	intermediate	21.910	63.648
5	1.000	age5r	1.000	intermediate	25.657	66.620
6	1.000	age6r	1.000	intermediate	MISSING	MISSING
7	2.000	age1r	1.000	low	29.311	56.000
8	2.000	age2r	1.000	low	22.690	60.000
9	2.000	age3r	1.000	low	24.199	62.133
10	2.000	age4r	1.000	low	21.403	65.859
11	2.000	age5r	1.000	low	17.220	67.565
12	2.000	age6r	1.000	low	22.994	70.953

Notice that the occasion variable **occ** has value labels **age1r**, **age2r**, **age3r, age4r, age5r** and **age6r**. The reason the variable has value labels is because it is defined as a categorical variable (you can see this in the **Names** window in the above screenshot as in the **categorical** column the variable **occ** is set to **True**). We shall change this variable to a numerical variable as this will remove the current value labels revealing the underling numeric values 1,2,3,4,5,6 which are the corresponding measurement occasions:

- Select the variable **occ** in the **Names** window and click the **Toggle Categorical button** to tell MLwiN that the variable is continuous (i.e. **categorical** = **False**)
- Click the **View** button again and you will see that the underlying numeric values


Data

goto line 1 | view | Help | Font | Show value labels

	id(26562)	occ(26562)	female(26562)	grade(26562)	phf(26562)	age(26562)
1	1.000	1.000	1.000	intermediate	39.591	55.000
2	1.000	2.000	1.000	intermediate	38.611	57.000
3	1.000	3.000	1.000	intermediate	39.928	60.416
4	1.000	4.000	1.000	intermediate	21.910	63.648
5	1.000	5.000	1.000	intermediate	25.657	66.620
6	1.000	6.000	1.000	intermediate	MISSING	MISSING
7	2.000	1.000	1.000	low	29.311	56.000
8	2.000	2.000	1.000	low	22.690	60.000
9	2.000	3.000	1.000	low	24.199	62.133
10	2.000	4.000	1.000	low	21.403	65.859
11	2.000	5.000	1.000	low	17.220	67.565
12	2.000	6.000	1.000	low	22.994	70.953

Two features of the data are immediately apparent:

- There is individual variation in the timing of measurements. For example, individual 1 is age 55 at occasion 1, while individual 2 is age 56.
- The length of time between measurements is not fixed and varies between individuals. For example, for individual 1, there is 2 years between occasions 1 and 2 and 3.42 years between occasions 2 and 3. The corresponding gaps for individual 2 are 4 and 2.13 years.

Later we will obtain summary statistics for the distributions of age and time between measurements across all individuals.

P15.1.3 Summarising longitudinal data

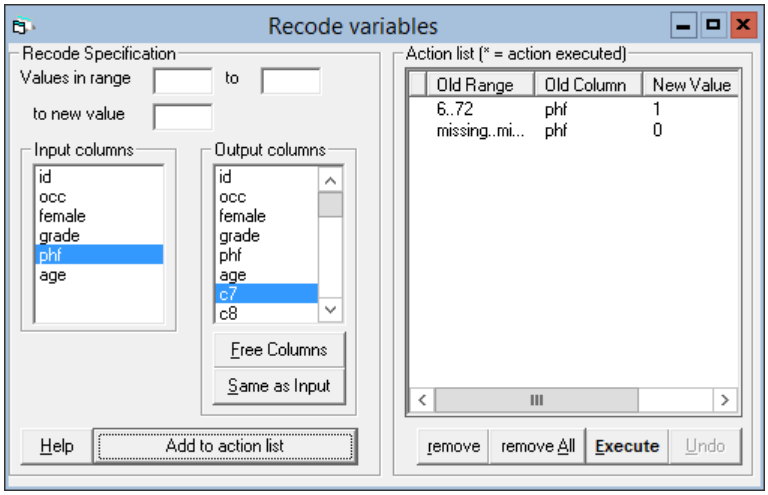
Before proceeding with growth curve analysis, we look at the extent of missing data and the distribution of the time between occasions.

Missing data patterns

We begin with a simple frequency table showing the number of valid (non-missing) values for our response **phf** at each occasion. In order to do this, we first generate a new binary indicator variable which equals 1 when **phf** is non-missing, and 0 otherwise.

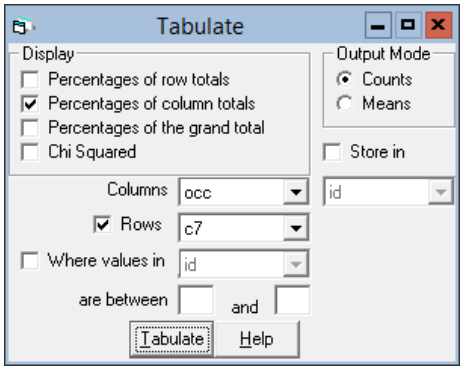
- From the **Data Manipulation menu**, select **recode**, then select **by Range**
- Under **Input columns**, select **phf** and under **Output columns** select **c7**. This tells the window to place the output of our data recoding into column **c7**.
- Enter **6** into the first **Values in range** box and **72** into the second **Values in range** box and **1** into the **to new value** box then click **Add to action list**. This ensures that any occasion at which the value of **phf** ranges from 6 to 72 will have a binary indicator value of 1. We chose the values 6 and 72 as these encompass the observed range of **phf** values as shown in the **Names** window (the min and max values of **phf** in the **Names** window are 6.40 and 71.86).

- Enter **missing** into the first **Values in range** box (type **m** and the box will automatically fill with 'missing'), **missing** into the second **Values in range** box and **0** into the **to new value** box then click **Add to action list**. This ensures that any occasion at which the value of **phf** is missing will have a binary indicator value of 0.
- Check that the window matches that shown below, then click **Execute**



Now we can tabulate our binary indicator for whether **phf** is missing.

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Percentages of column totals** box in order to obtain column percentages as well as frequencies in the subsequent cross-tabulation
- In the **Columns** drop-down box select **occ** to obtain the number of individuals at each occasion
- Check the **Rows** box and enter **c7** into the drop-down box to ensure that the frequencies and percentages are reported separately by our new binary indicator variable **c7** (i.e., by whether **phf** is missing)
- Check that the window matches that shown below, then click **Tabulate**

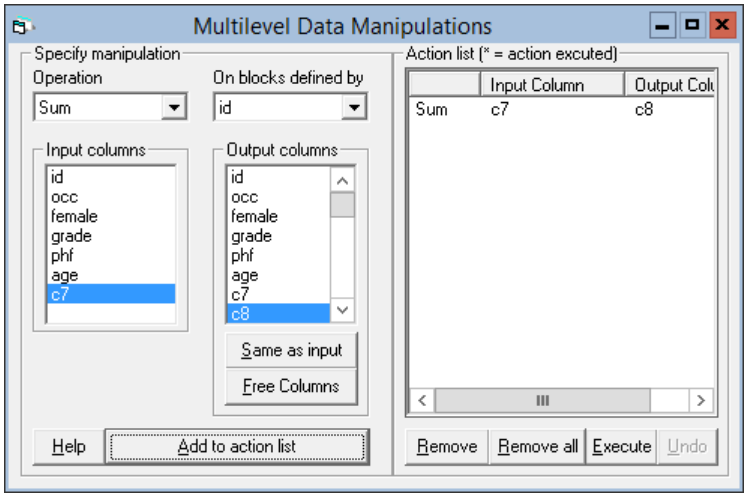


```
->TABUlate 2  'occ'   'c7'

Columns are levels of occ
Rows are levels of c7

                    1          2          3          4          5          6     TOTALS

  0    N          251        659       1024       1146       1176       1181       5437
  COL  %          5.7       14.9       23.1       25.9       26.6       26.7       20.5

  1    N         4176       3768       3403       3281       3251       3246      21125
  COL  %         94.3       85.1       76.9       74.1       73.4       73.3       79.5

 TOTALS          4427       4427       4427       4427       4427       4427      26562
  COL  %        100.0      100.0      100.0      100.0      100.0      100.0      100.0
```

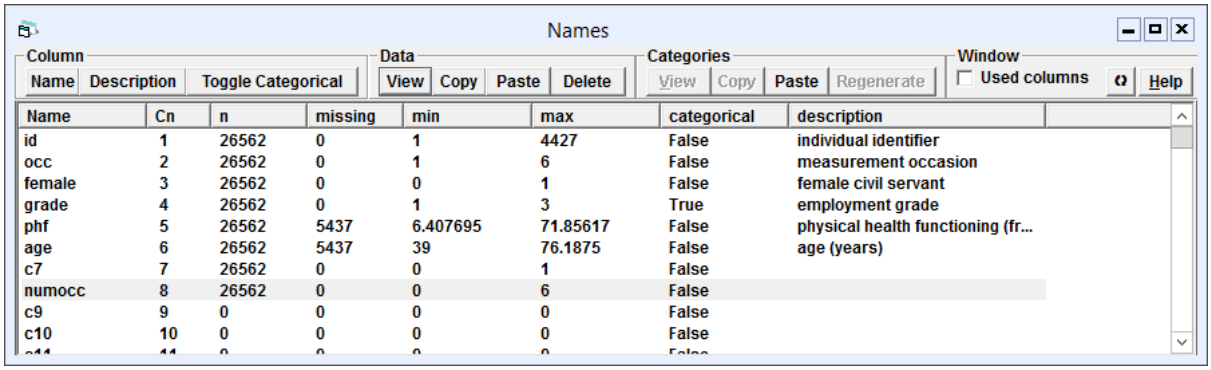
The total number of individuals in the dataset is 4427, of whom 4176 (94%) were present at occasion 1, falling to 3246 (73%) at occasion 6.

Next we obtain the number of non-missing observations per individual by using the **Multilevel data manipulations** window.

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- In the **Operation** box select **Sum** (i.e., summation)
- In the **On blocks defined by** box select **id** to ensure that the summation operation is conducted for each individual in the dataset separately
- Select **c7** from **Input columns** (so that the summation will be over the values of the binary indicator for whether **phf** is observed or missing for each individual)
- Select **c8** from **Output columns**
- Click **Add to action list**,
- Check that the window matches that shown below, then click **Execute**
- From the **Data Manipulation** menu, select **Command interface** and run the command **NAME c8 'numocc'** to rename the new variable **c8** to **numocc**



The **Names** window should now look like this:



Name	Cn	n	missing	min	max	categorical	description
id	1	26562	0	1	4427	False	individual identifier
occ	2	26562	0	1	6	False	measurement occasion
female	3	26562	0	0	1	False	female civil servant
grade	4	26562	0	1	3	True	employment grade
phf	5	26562	5437	6.407695	71.85617	False	physical health functioning (fr...
age	6	26562	5437	39	76.1875	False	age (years)
c7	7	26562	0	0	1	False	
numocc	8	26562	0	0	6	False	
c9	9	0	0	0	0	False	
c10	10	0	0	0	0	False	

We can now tabulate the **numocc** variable to obtain a breakdown of the number of missing observations in the dataset as a whole at each time point.

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Percentages of column totals** box to tell MLwiN to return percentages alongside column totals
- In the **Columns** drop-down box select **numocc**
- Click **Tabulate**

```
->TABUlate 2  'numocc'

                    0          1          2          3          4          5          6     TOTALS

       N          438       1464       1896       1848       2490       4566      13860      26562
       %          1.6        5.5        7.1        7.0        9.4       17.2       52.2      100.0
```

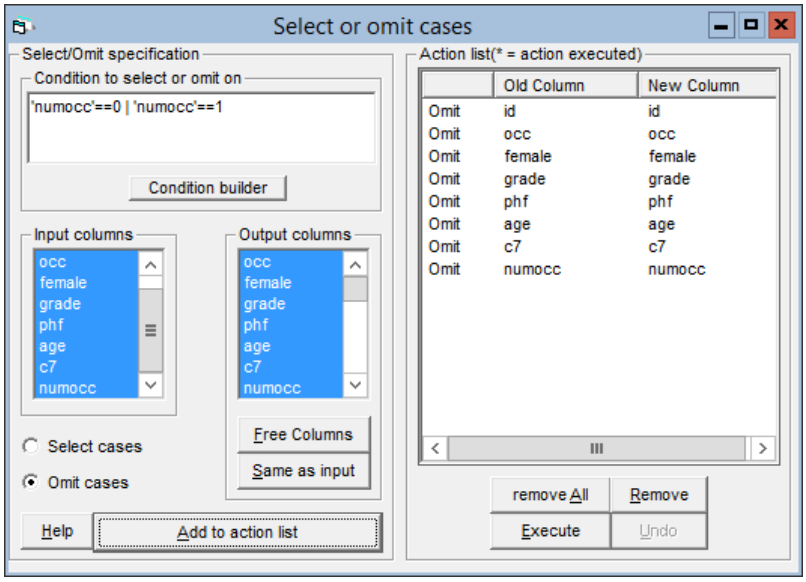
The tabulation of **numocc** is across all observations and as such each individual appears six times (there are six time points per individual hence the total number of 26562 observations for the 4427 individuals). To save dividing all of the **numocc** categories by hand, we can recover the correct values by simply performing a two-way tabulation of **numocc** by **occ**.

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Rows** box and select **occ** in the drop-down box to create a two-way tabulation between **numocc** and **occ**
- Click **Tabulate**

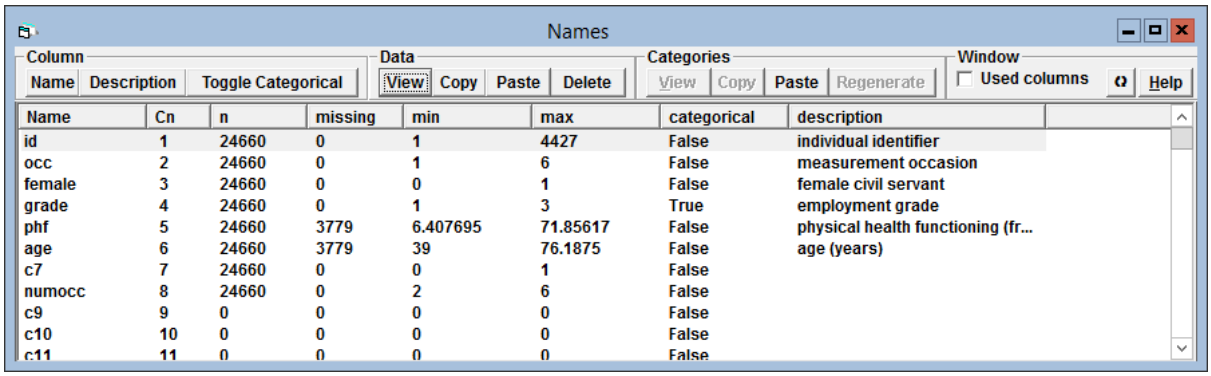
```
->TABUlate 2  'numocc'   'occ'

Columns are levels of numocc
Rows are levels of occ


                    0          1          2          3          4          5          6     TOTALS

   1   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7

   2   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7

   3   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7

   4   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7

   5   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7

   6   N           73        244        316        308        415        761       2310       4427
   COL %         16.7       16.7       16.7       16.7       16.7       16.7       16.7       16.7


  TOTALS          438       1464       1896       1848       2490       4566      13860      26562
   COL %        100.0      100.0      100.0      100.0      100.0      100.0      100.0      100.0
```

We now have a tabulation in which we can quickly see that a total of 317 (= 73 + 244) individuals have either completely missing data or only one valid response. The individuals with **numocc** = 0 will automatically be deleted from any analysis, but we will also delete those with only 1 record because they contribute no information about change in the response.

- From the **Data Manipulation** menu, select **Select or omit observations**
- Under **Condition to select or omit on**, click the **condition builder** button and in the new window which pops up type **'numocc'==0 | 'numocc'==1** into the box in the top right corner of the window and then click **Done**. The vertical bar | in the above condition is the logical "or" operator
- Select **all** of the variables in the **Input columns** box, as we want to delete all individuals with 0 or only 1 observed **phf** scores
- Select **Omit observations** to tell MLwiN to drop all records where the condition is true
- Click **Same as input** under the **Output columns** in order to write over the existing data
- Click **Add to action list**
- Check that the window matched that shown below, then click **Execute**



The **Names** window should now look as follows. The number of observations has reduced to 24660 (from 26562) after a total of 6 × 317 = 1902 records have been dropped from the full dataset.



Name	Cn	n	missing	min	max	categorical	description
id	1	24660	0	1	4427	False	individual identifier
occ	2	24660	0	1	6	False	measurement occasion
female	3	24660	0	0	1	False	female civil servant
grade	4	24660	0	1	3	True	employment grade
phf	5	24660	3779	6.407695	71.85617	False	physical health functioning (fr...
age	6	24660	3779	39	76.1875	False	age (years)
c7	7	24660	0	0	1	False	
numocc	8	24660	0	2	6	False	
c9	9	0	0	0	0	False	
c10	10	0	0	0	0	False	
c11	11	0	0	0	0	False	

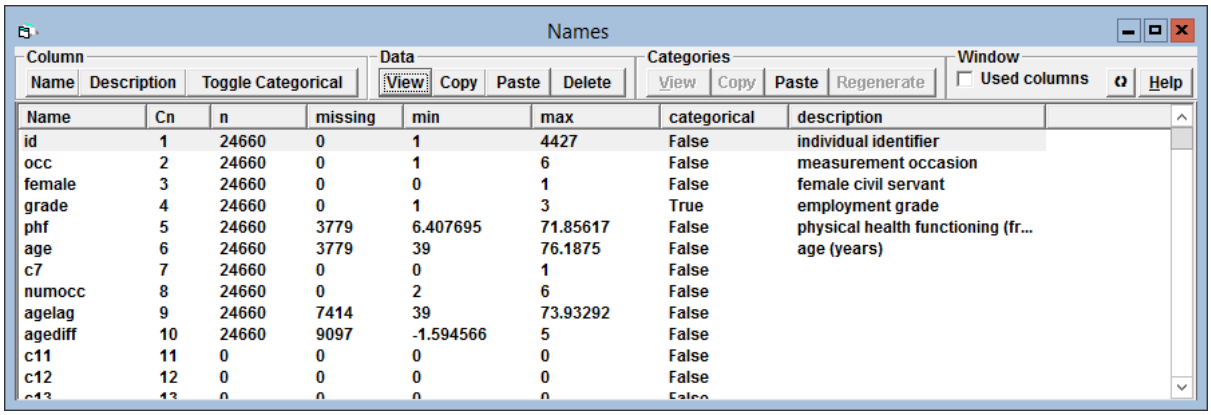
Time between occasions

In order to examine the length of time between occasions, we must first create a new variable called **agelag** which equals the age at the previous occasion (i.e., the first order lag of age), and then we can create a new variable called **agediff** which equals the difference in age between adjacent occasions.

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- In the **Operation** box select **Lags**
- In the **On blocks defined by** box select **id** to ensure that lags are only created within unique individuals
- Select **age** from **Input columns**
- Select **c9** from **Output columns**
- Click **Add to action list**, and then **Execute**
- From the **Data Manipulation** menu, select **Command interface** and run the command **NAME c9 'agelag'** to name this new variable **agelag**

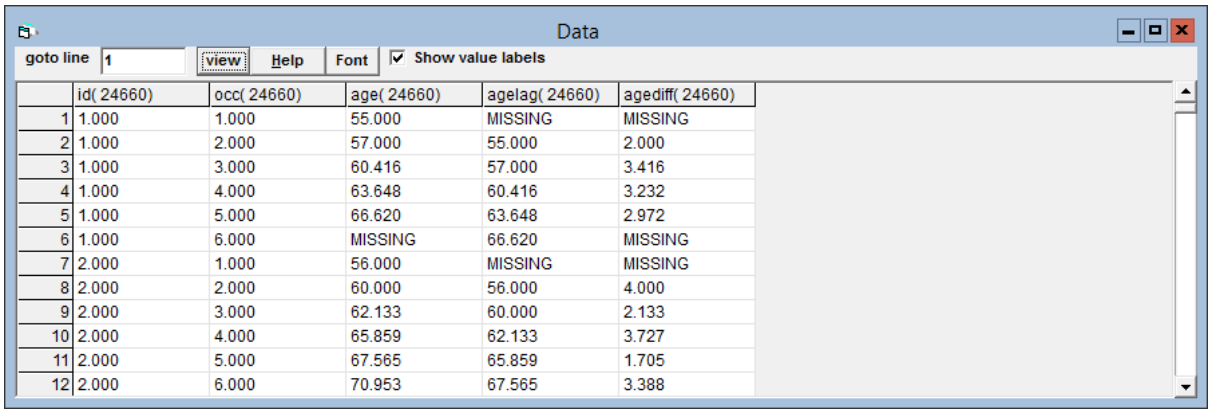
- Run the command **CALC c10 = 'age' - 'agelag'** to create a new variable in **c10** that contains at each occasion the difference in age between the current and previous occasion
- Run the command **NAME c10 'agediff'** to name this new variable **agediff**
- In the **Names** window, select the variables **id**, **occ**, **age**, **agelag** and **agediff** then click the **View** button to view the new data

The **Names** window should look like this:



Name	Cn	n	missing	min	max	categorical	description
id	1	24660	0	1	4427	False	individual identifier
occ	2	24660	0	1	6	False	measurement occasion
female	3	24660	0	0	1	False	female civil servant
grade	4	24660	0	1	3	True	employment grade
phf	5	24660	3779	6.407695	71.85617	False	physical health functioning (fr...
age	6	24660	3779	39	76.1875	False	age (years)
c7	7	24660	0	0	1	False	
numocc	8	24660	0	2	6	False	
agelag	9	24660	7414	39	73.93292	False	
agediff	10	24660	9097	-1.594566	5	False	
c11	11	0	0	0	0	False	
c12	12	0	0	0	0	False	

In the **Data** window your data should now look as follows:



	id(24660)	occ(24660)	age(24660)	agelag(24660)	agediff(24660)
1	1.000	1.000	55.000	MISSING	MISSING
2	1.000	2.000	57.000	55.000	2.000
3	1.000	3.000	60.416	57.000	3.416
4	1.000	4.000	63.648	60.416	3.232
5	1.000	5.000	66.620	63.648	2.972
6	1.000	6.000	MISSING	66.620	MISSING
7	2.000	1.000	56.000	MISSING	MISSING
8	2.000	2.000	60.000	56.000	4.000
9	2.000	3.000	62.133	60.000	2.133
10	2.000	4.000	65.859	62.133	3.727
11	2.000	5.000	67.565	65.859	1.705
12	2.000	6.000	70.953	67.565	3.388

We now calculate summary statistics for **agediff** for each occasion as follows:

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Means** tick in the **Output Mode** box
- Select **agediff** in the **Variate column** box to obtain summary statistics of **agediff**
- Select **occ** in the **Columns** box to calculate summary statistics separately at each occasion
- Click **Tabulate**

The output window should display the number of individuals at each occasion along with the mean and standard deviation of the variable **agediff**.

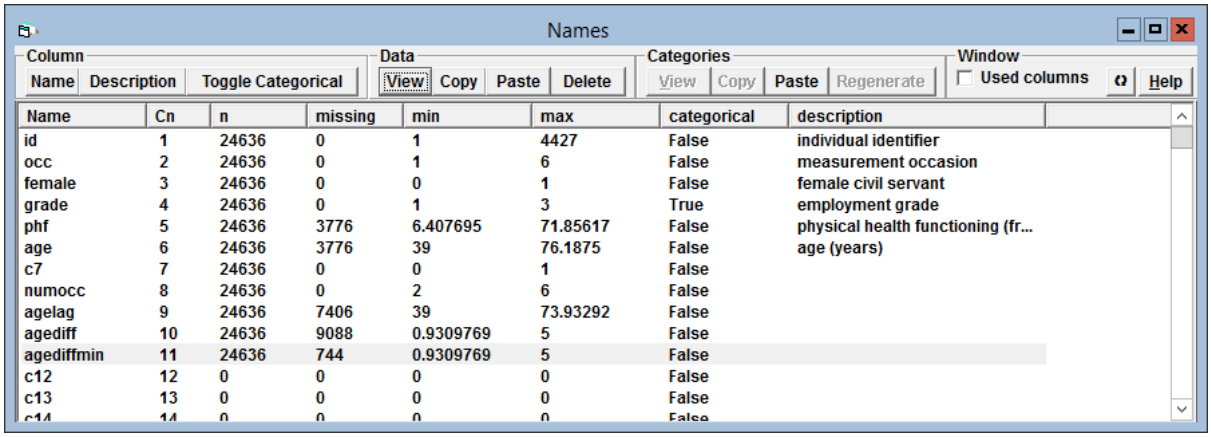
```
->TABUlate  'agediff'    'occ'

9097 missing value(s)


Variable tabulated is agediff

              1          2          3          4          5          6     TOTALS

      N       0       3625       3154       2941       2904       2939      15563
  MEANS       *      3.035      3.245      2.994      2.464      2.914      2.940
   SD'S       *      0.624      0.538      0.454      0.490      0.347      0.565
```

The mean age difference between consecutive occasions ranges from around 2.5 years to just over 3 years, but there is substantial variation between individuals.

However, if you look at the **Names** window shown above we see that there are some negative age differences. These are not sensible and may reflect data errors. We will identify individuals with an age difference that is zero or negative and remove them from the dataset. We next calculate the minimum value of **agediff** for each individual (**agediffmin**), and drop observations for which this variable is less than or equal to zero.

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- In the **Operation** box select **Minimum** so that MLwiN knows to select the minimum value
- In the **On blocks defined by** box select **id** so that the data manipulation is carried out separately for each individual
- Select **agediff** from **Input columns**
- Select **c11** from **Output columns**
- Click **Add to action list**, and then **Execute**
- From the **Data Manipulation** menu, select **Command interface** and run the command **NAME c11 'agediffmin'** to rename this new variable
- From the **Data Manipulation** menu, select **Select or omit observations**
- Click in the **Condition to select or omit on box** and enter **'agediffmin'<=0** so that MLwiN knows to omit all observations where the minimum age difference for an individual is zero or negative
- Select all of the variables listed in the **Input columns** box
- Select **Omit observations**
- Click **Same as input** under the **Output columns**
- Click **Add to action list**, and then click **Execute**

In the **Names** windows you will now see that the minimum value of **agediff** is positive and that 24 observations have been removed from the dataset (the sample sized has reduced from 24660 to 24636).



Name	Cn	n	missing	min	max	categorical	description
id	1	24636	0	1	4427	False	individual identifier
occ	2	24636	0	1	6	False	measurement occasion
female	3	24636	0	0	1	False	female civil servant
grade	4	24636	0	1	3	True	employment grade
phf	5	24636	3776	6.407695	71.85617	False	physical health functioning (fr...
age	6	24636	3776	39	76.1875	False	age (years)
c7	7	24636	0	0	1	False	
numocc	8	24636	0	2	6	False	
agelag	9	24636	7406	39	73.93292	False	
agediff	10	24636	9088	0.9309769	5	False	
agediffmin	11	24636	744	0.9309769	5	False	
c12	12	0	0	0	0	False	
c13	13	0	0	0	0	False	
c14	14	0	0	0	0	False	

PART I: GROWTH CURVE MODELS

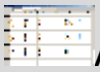
P15.2 Introduction to Growth Curve Models

There is no practical for this section.

P15.3 Linear Growth Model for Continuous Repeated Measures

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to  ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "  15.3.wsz"

P15.3.1 Preliminary analysis of the physical functioning dataset

Before fitting any growth models, we carry out some further exploratory analysis of the data. We begin by calculating descriptive statistics for the physical functioning response and age at each occasion.

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Means** tick in the **Output Mode** box to calculate sample means
- Select **phf** in the **Variate column** box
- Select **occ** in the **Columns** box
- Click **Tabulate**

The resulting table reports the mean and standard deviation of **phf** separately at each occasion as well as the number of individuals for which **phf** is observed.

```
->TABUlate  'phf'    'occ'

Variable tabulated is phf

              1          2          3          4          5          6     TOTALS

      N    3971       3735       3397       3273       3246       3238      20860
  MEANS     52.359     50.737     51.034     50.242     48.857     48.882     50.436
   SD'S      7.005      8.227      8.050      8.615      9.054      9.361      8.458
```

There is some suggestion of a decline in mean functioning and an increase in between-individual variance with occasion.

- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Means** tick in the **Output Mode** box
- Select **age** in the **Variate column** box
- Select **occ** in the **Columns** box
- Click **Tabulate**

```
->TABUlate  'age'    'occ'

Variable tabulated is age

              1          2          3          4          5          6     TOTALS

      N    3971       3735       3397       3273       3246       3238      20860
  MEANS     49.528     52.647     55.784     58.860     61.124     63.906     56.605
   SD'S      6.049      6.044      6.000      6.018      5.976      5.939      7.796
```

The mean age increases, as expected, but there is substantial variation in age at each occasion. We can also see that the gap in mean age between consecutive measurements differs across occasions (as seen directly in P15.1.3 when we calculated age differences and examined summary statistics for each occasion).

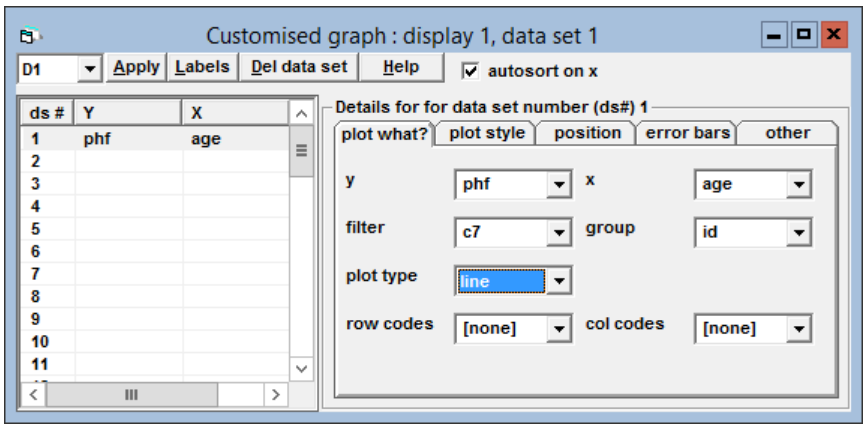
Because of the variation in age at any given occasion, we will obtain different results according to whether age or occasion is used as the time variable. We therefore have to choose the most appropriate time metric to use in further analysis. This decision did not arise in the reading example of the Concepts part of this module because, although there was some variation in age at each measurement occasion, the individual ages were not available. In the present case, age would seem the more relevant measure of time. We would expect a person's physical functioning to depend on their age, while whether a particular measurement was taken at wave 1 or 3 in the study is unimportant to the respondent. We will therefore use age as the time variable in most of the growth curve analysis that follows.[3]

Our final descriptive analysis will be a plot of physical functioning trajectories by age for the first 5 individuals in the dataset. To produce this plot, we need to first create a new binary indicator variable identifying the first 5 individuals before we can plot their trajectories using **Customised Graph(s)**. We do this by generating a new variable which is equal to the individual identifier **id**, but where we recode values in the range 1 to 5 to 1 and values above 5 to 0.

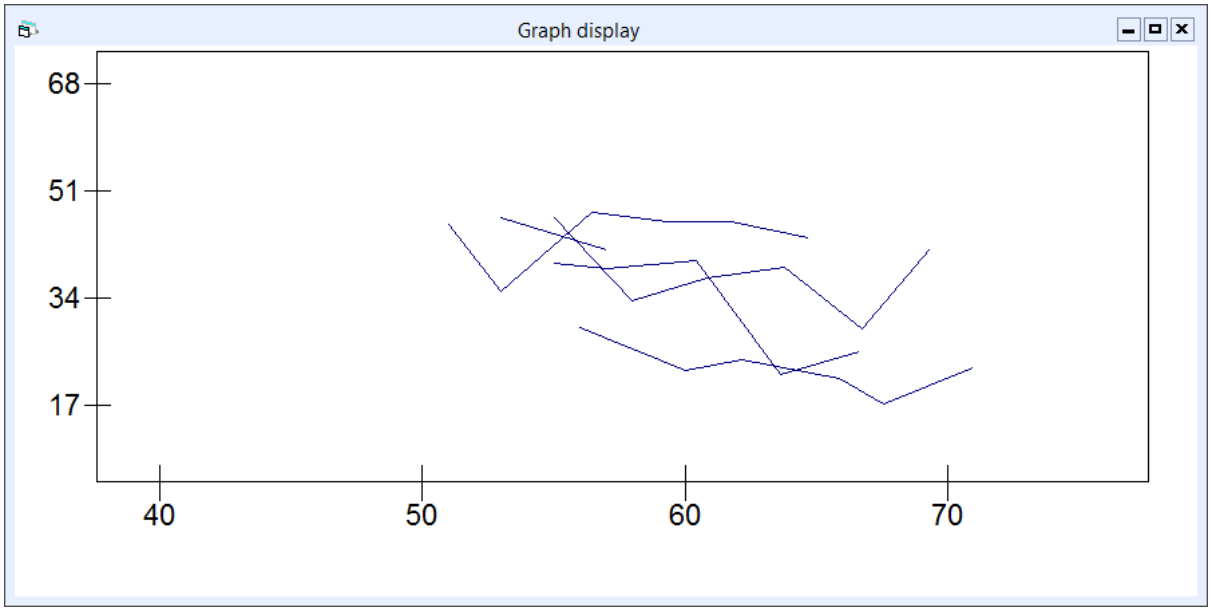
- From the **Data Manipulation** menu, select **recode** menu, select **by Range**
- In **Input columns** select **id** and in **Output columns** select **c7**
- Enter **1** into the first **Values in range** box, **5** into the second **Values in range** box and **1** into the **to new value** box then click **Add to action list**. This tells MLwiN to give the first five individuals (values 1 to 5) a value of 1 in our new binary indicator variable

[3] In some of the subsequent analyses, we will use occasion but only after restricting the sample so that there is little variation in age between individuals at any given occasion.

- Enter **6** into the first **Values in range** box, **4427** into the second **Values in range** box and **0** into the **to new value** box then click **Add to action list**. This tells MLwiN to give all other individuals (values 6 to 4427) a value of 0 in our new binary indicator variable
- Click **Execute**
- Now that we have our indicator variable created in **c7**, we can plot our graph
- From the **Graphs** menu, select **Customised Graph(s)**
- In the **plot what?** tab select **phf** as the **y** variable and **age** as the **x** variable, select the new binary indicator variable **c7** as the filter variable (so that MLwiN 'filters' out all but the first five individuals in the dataset), **id** as the group variable so that the observed values of **phf** are connected separately for each individual, change **plot type** to **line**
- Check that the window matches that shown below, then click **apply**



You should see the following graph which plots **phf** scores on the *y*-axis against **age** on the *x*-axis. There are five lines plots, one for each of the first five individuals in the data. Axis titles can be edited and various other changes to the graph can be made by right clicking on the graph, but we leave this as an exercise to the reader.



The trajectories start at different points because of the variation in age at the first occasion. It is also apparent that individuals vary in the number of observations they contribute. Although the trajectories for these individuals are clearly nonlinear, there is some suggestion of a decline in physical functioning with age.

P15.3.2 Random intercept model

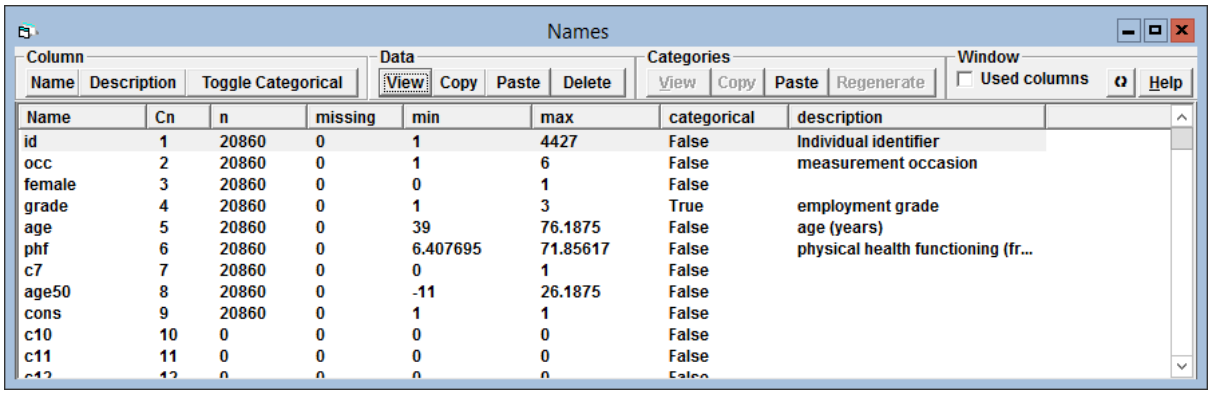
We are almost ready to fit our first growth curve model, but we will first centre the age variable to aid interpretation of the model estimates. This is especially important when we move to random slope models. We will centre around 50 years which is close to the mean age at the first occasion. The new variable **age50** will be used in place of age in the subsequent analysis.

- From the **Data Manipulation** menu, select **Command interface**
- Type **CALC c8 = 'age' - 50** to create a new variable for age that is centred around the value of 50
- Type **NAME c8 'age50'** into the **Command interface** window to rename this new variable

We also need to generate a new variable named **cons** which contains a series of 1's that MLwiN uses as the variable associated with the intercept in its models.

- From the **Data Manipulation** menu, select **Generate vector**
- In **Output column** select **c9**
- In **Number of copies** enter **20860** to ensure that the new variable is created for each of the 20860 observations in the dataset
- In **Value** enter **1**
- Click **Generate**
- From the **Data Manipulation** menu, select **Command interface**
- Type **NAME c9 'cons'** to rename the variable

The **Names** window should now look like this



Name	Cn	n	missing	min	max	categorical	description
id	1	20860	0	1	4427	False	Individual identifier
occ	2	20860	0	1	6	False	measurement occasion
female	3	20860	0	0	1	False	
grade	4	20860	0	1	3	True	employment grade
age	5	20860	0	39	76.1875	False	age (years)
phf	6	20860	0	6.407695	71.85617	False	physical health functioning (fr...
c7	7	20860	0	0	1	False	
age50	8	20860	0	-11	26.1875	False	
cons	9	20860	0	1	1	False	
c10	10	0	0	0	0	False	
c11	11	0	0	0	0	False	

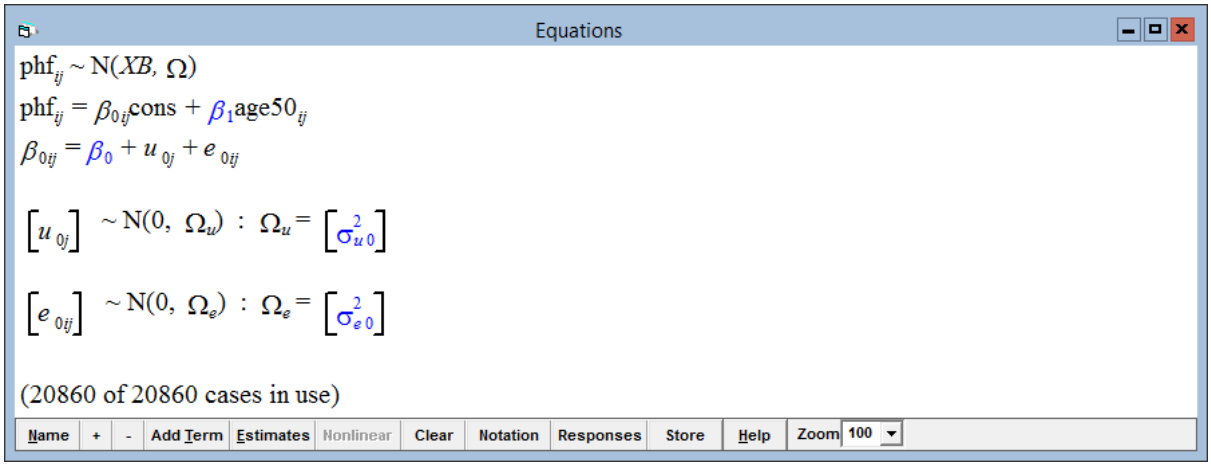
We begin by fitting a random intercept model with **age50** as the only predictor. The model takes the form of equation (15.1) in C15.3.2:

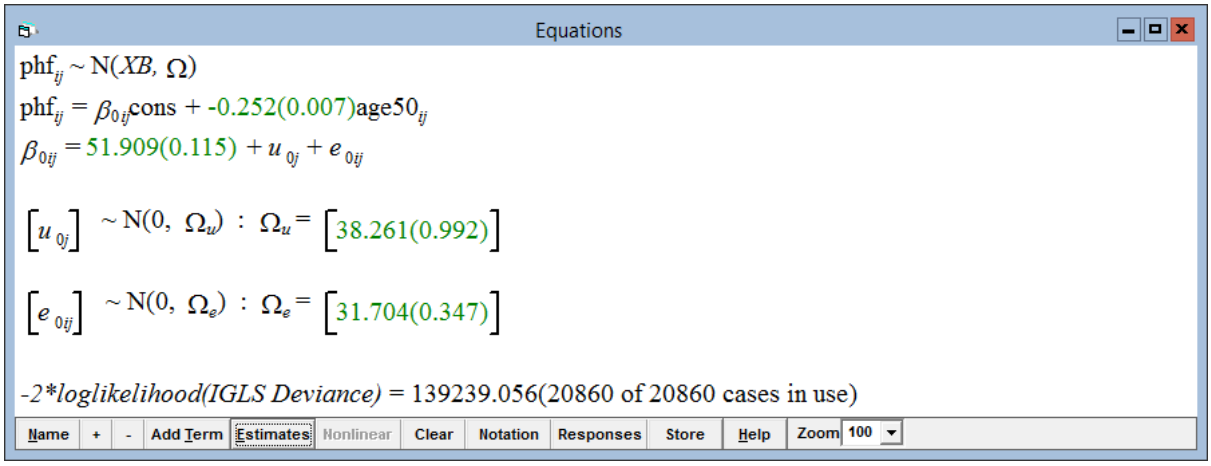
$$\text{phf}_{ij} = \beta_{0j} + \beta_1 \text{age50}_{ij} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$

where $u_{0j} \sim N(0, \sigma^2_{u0})$ and $e_{ij} \sim N(0, \sigma^2_e)$.

As usual, we specify and fit all models using the **Equations** window.

- From the **Model** menu, select **Equations**
- To specify **phf** as the outcome variable in the model, click on the red **y** and select **phf** as the **y** variable, then **2-ij** as the **N** levels variable to specify a two-level model, **id** as the **level 2(j)** variable, and **occ** as the **level 1(i)** variable. This model specification respects the nesting of occasions within individuals
- To add the constant to the model, click on **Add Term** and select **cons** in the variable drop-down box then click **Done**. You should see that a new term for cons ($\beta_0 cons$) has appeared in the **Equations** window
- Click on the new term β_0 and check the **j(id)** and **i(occ)** boxes to allow the cons term to vary at the occasion (**i**) and individual (**j**) level then click **done**. This introduces the residual and random intercept effect into the model.
- Click on **Add Term** and select **age50** in the variable drop-down box to add the centred age term then click **Done**
- Click **Estimates** to show the full model specification (which should look like the first screenshot below). The four parameters to be estimated are highlight in blue
- Click **Start** to run the model
- Once the model has converged, the parameters turn from blue to green. Click the **Estimates** button a second time to reveal the parameter estimates and standard errors (displayed in parentheses)





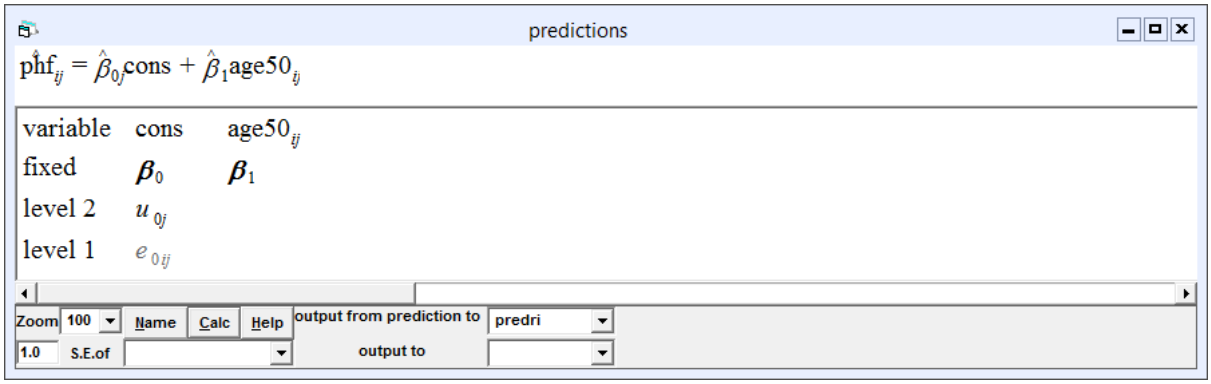
The fitted mean linear trajectory has equation:

$$\widehat{\text{phf}}_{ij} = 51.91 - 0.25\ \text{age50}_{ij}$$

The mean predicted value of **phf** is 51.91 at age 50 and declines significantly with age. The between-individual variance in physical functioning is estimated as 38.26, and the within-individual variance estimated as 31.70. After adjusting for the linear age trend, the correlation between any pair of **phf** values for the same individual is estimated as 38.26/(38.26 + 31.70) = 0.55. Equivalently, we can say that 55% of the total residual variance in physical functioning is due to differences between individuals.

To gain an impression of how well the random intercept model fits the data, we can superimpose a plot of the predicted trajectories from the model on a plot of the observed values of **phf**. We first compute predicted values of **phf** for each individual by age using the **Predictions** window.

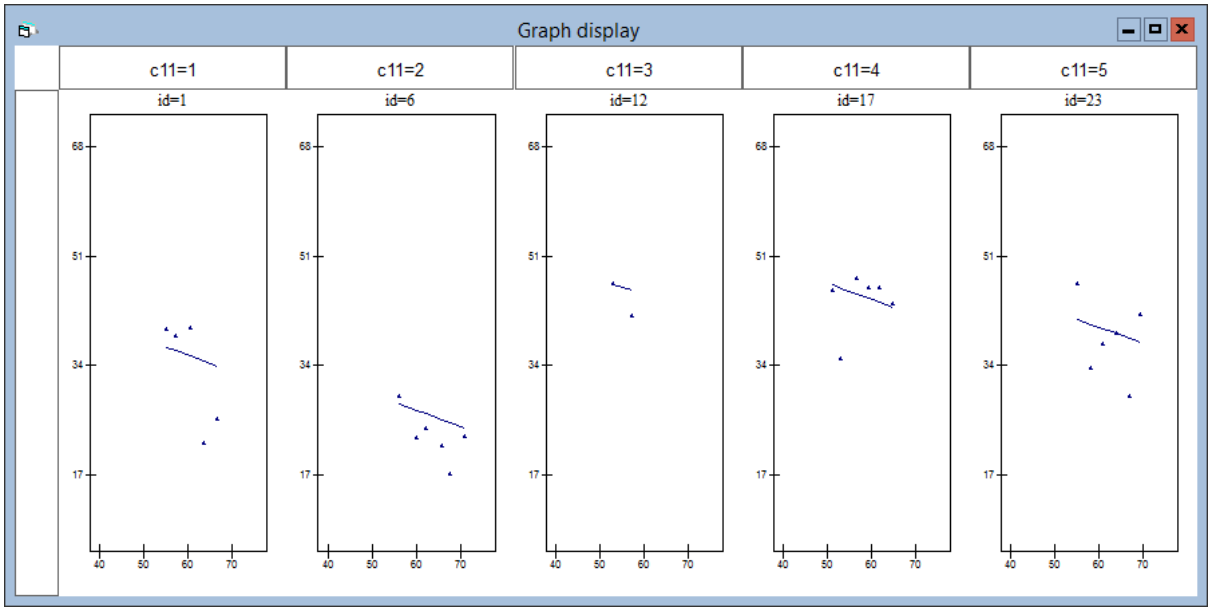
- From the **Model** menu, select **Predictions**
- Click on $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{u}_{0j}$ at the bottom of the window to select these terms (they will turn from grey to black as they are selected)
- In the **output from prediction to** drop-down box select **c10**
- Press **Ctrl+N** to bring up the **Rename** window and type **predri** and click **Rename** to rename **c10** to **predri**
- Check that the **Predictions** window matches that shown below, then click **Calc**



The predictions are stored in a new variable called **predri.** We will now plot the observed values of **phf** against **age** as points, and the predicted values **predri** against **age** as lines (i.e., trajectories or growth curves). We shall again restrict this plot to the first five individuals in the data.

- From the **Data Manipulation** menu, select **recode** then select **by Range**
- In **Input columns** select **id** and in **Output columns** select **c11**
- Enter **6** into the first **Values in range** box, **4427** into the second **Values in range** box and **missing** into the **to new value** box then click **Add to action list**, and then **Execute**
- From the **Graphs** menu, select **Customised Graph(s)**
- On the **plot what?** tab select **point** from the plot type drop-down box, and in the **col codes** box select **c11**. This variable equals the original individual identifier **id** for the first five individuals in the data, but is missing for all other individuals. It is by using this variable as the columns of the resulting figure that we restrict the graph to the first five individuals in the data.
- Click on the second line of the left hand window (**ds#**) so that MLwiN knows to create and overlay two separate graphs in the same plot
- In the **plot what?** tab select **predri** as the **y** variable and **age** as the **x** variable, change **plot type** to **line**
- Click **apply**

After a short wait you should see the following graph:



We can see that, at least for these five individuals, the random intercept model does not appear to fit the observed data very well. The assumptions that the fitted individual trajectories are parallel and linear are overly restrictive. We will start by relaxing the first of these restrictions by allowing the slope of age to vary across individuals.

In the analysis of the reading dataset in C15.3.2 the within-individual correlation matrix implied by the fitted model was compared with the sample correlation matrix. In that example, there was a single correlation matrix because individuals were measured at fixed occasions, i.e. $t_{ij} = t_i$. When there is individual variation in the timing of measurements, however, there is no longer a single correlation matrix because the correlations will depend on t_{ij}. In the physical functioning dataset, for example, we would expect the correlation between two measurements of **phf** on an individual to depend on the age at which the measurements were taken. If each individual has a unique set of values for age across the measurement occasions, a different correlation matrix is defined for each individual. In the following analyses, we compute the estimated correlation matrix implied by different models for the first two individuals. This is done to illustrate differences in model assumptions, rather than to assess model fit.

We can obtain the model-implied within-individual correlations and standard deviations for the first two individuals in the dataset using the following sequence of commands.

- From the **Data Manipulation** menu, select **Command interface**
- Click on **Output**
- Enter the following commands one block at a time and check that your output matches that shown below:

  VMAT 1 c100

  CALC g1 = c100

  PRIN g1
- **g1** in the output window shows the model-implied covariance matrix for individual 1 (**VMAT 1** tells MLwiN to use the first individual/cluster in the dataset)

  CALC g2 = SQRT(DIAG(g1))

  PRIN g2
- **g2** in the output window shows the model-implied standard deviations for individual 1 (i.e. the square-root **SQRT** of the diagonal elements **DIAG** of the model-implied covariance matrix **g1**)

  CALC g3 = 1/g2

  CALC g4 = ~(g3*g1)*g3

  PRIN g4
- **g4** in the output window shows the model-implied correlation matrix for individual 1 (i.e., the covariance matrix pre and post multiplied by inverse diagonal matrices holding the standard deviation)

  VMAT 2 c101

  CALC g5 = c101

  PRIN g5
- **g5** in the output window shows the model-implied covariance matrix for individual 2

  CALC g6 = SQRT(DIAG(g5))

PRIN g6

- **g6** in the output window shows the model-implied standard deviations for individual 2

CALC g7 = 1/g6

CALC g8 = ~(g7*g5)*g7

PRIN g8

- **g8** in the output window shows the model-implied correlation matrix for individual 2

The standard deviations and correlations for individuals 1 and 2 are shown below. Individual 1 has 5 measurements while individual 2 has 6, but the estimated standard deviations and correlations are the same, as expected for a random intercept model. Moreover the standard deviation is constant across occasions. The standard deviation is equal to the square root of the total variance ($\sqrt{38.26 + 31.70}$), and the correlation is the ratio of the between-individual variance to the total variance (as computed earlier).

```
->VMAT 1 c100
->CALC g1 = c100
->PRIN g1

      c1496         c1497         c1498         c1499         c1500
 N =     5             5             5             5             5
   1  69.965        38.261        38.261        38.261        38.261
   2  38.261        69.965        38.261        38.261        38.261
   3  38.261        38.261        69.965        38.261        38.261
   4  38.261        38.261        38.261        69.965        38.261
   5  38.261        38.261        38.261        38.261        69.965
```

```
->CALC g2 = SQRT(DIAG(g1))
->PRIN g2

      c1495
 N =     5
   1  8.3645
   2  8.3645
   3  8.3645
   4  8.3645
   5  8.3645
```

```
->CALC g3 = 1/g2
->CALC g4 = ~(g3*g1)*g3
->PRIN g4

      c1489         c1490         c1491         c1492         c1493
 N =     5             5             5             5             5
   1  1.0000        0.54686       0.54686       0.54686       0.54686
   2  0.54686       1.0000        0.54686       0.54686       0.54686
   3  0.54686       0.54686       1.0000        0.54686       0.54686
   4  0.54686       0.54686       0.54686       1.0000        0.54686
   5  0.54686       0.54686       0.54686       0.54686       1.0000
```

```
->VMAT 2 c101
->CALC g5 = c101
->PRIN g5

      c1483         c1484         c1485         c1486         c1487         c1488
 N =      6             6             6             6             6             6
   1  69.965        38.261        38.261        38.261        38.261        38.261
   2  38.261        69.965        38.261        38.261        38.261        38.261
   3  38.261        38.261        69.965        38.261        38.261        38.261
   4  38.261        38.261        38.261        69.965        38.261        38.261
   5  38.261        38.261        38.261        38.261        69.965        38.261
   6  38.261        38.261        38.261        38.261        38.261        69.965
```

```
->CALC g6 = SQRT(DIAG(g5))
->PRIN g6

      c1482
 N =      6
   1  8.3645
   2  8.3645
   3  8.3645
   4  8.3645
   5  8.3645
```

```
->CALC g7 = 1/g6
->CALC g8 = ~(g7*g5)*g7
->PRIN g8

      c1475         c1476         c1477         c1478         c1479         c1480
 N =      6             6             6             6             6             6
   1  1.0000        0.54686       0.54686       0.54686       0.54686       0.54686
   2  0.54686       1.0000        0.54686       0.54686       0.54686       0.54686
   3  0.54686       0.54686       1.0000        0.54686       0.54686       0.54686
   4  0.54686       0.54686       0.54686       1.0000        0.54686       0.54686
   5  0.54686       0.54686       0.54686       0.54686       1.0000        0.54686
   6  0.54686       0.54686       0.54686       0.54686       0.54686       1.0000
```

P15.3.3 Random slope model

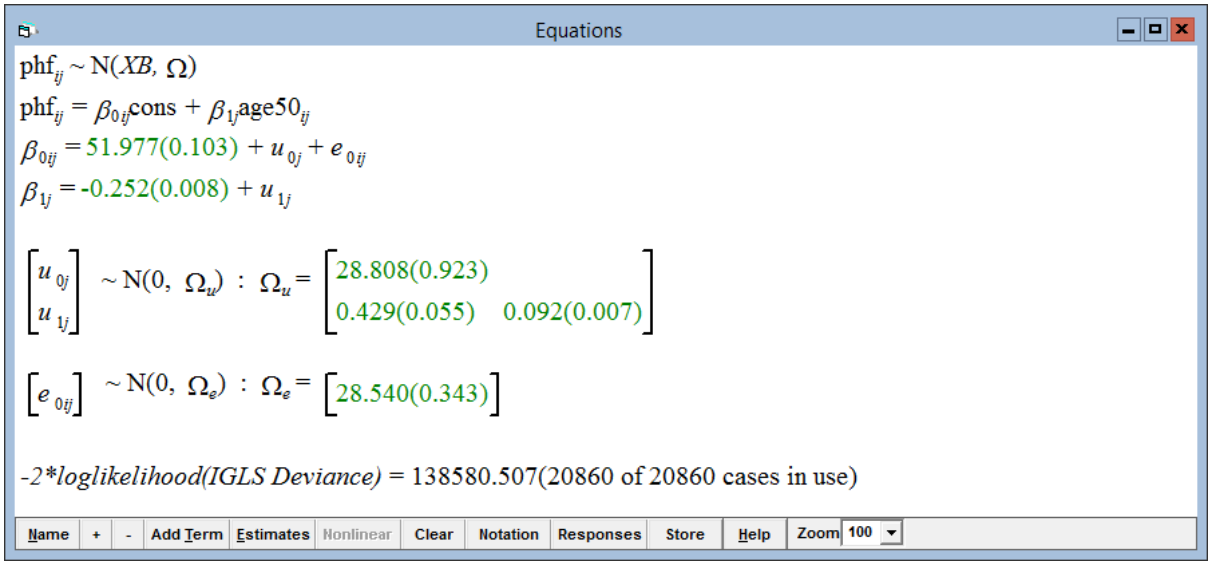
We now relax the restriction that all individuals have the same slope for age by allowing the coefficient of **age50** to vary across individuals. In doing so, we also allow the variance in functioning and the within-individual correlation to depend on age. The model takes the form of equation (15.2) in C15.3.3:

$$\begin{aligned}\text{phf}_{ij} &= \beta_{0j} + \beta_{1j}\text{age50}_{ij} + e_{ij}\\ \beta_{0j} &= \beta_0 + u_{0j}\\ \beta_{1j} &= \beta_1 + u_{1j}\end{aligned}$$

$e_{ij} \sim N(0, \sigma_e^2)$ and $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_u)$ where $\Omega_u = \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix}$

We can specify and fit this model in the **Equations** model as follows.

- In the **Equations** window click on **age50** to open the **X variable** window and check the **j(id)** box to include a random slope for age, then click **Done**
- Click **Start** to run the model



There is little change in the fixed part estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$). The intercept variance σ^2_{u0} is estimated as 28.81 which is the between-individual variance at age 50. The slope variance σ^2_{u1} is estimated as 0.09 which is the between-individual variance in growth rates. The estimated intercept-slope covariance σ_{u01} is 0.43 and the correlation is $0.43/\sqrt{28.81 \times 0.09} = 0.27$. A positive covariance combined with the negative estimate of β_1 suggests that individuals with above-average functioning at age 50 ($u_{0j} > 0$ and $\beta_0 + u_{0j} > \beta_0$) will tend to have below-average slopes ($u_{1j} > 0$ and $\beta_1 + u_{1j} < \beta_1$).[4]

Compared to the random intercept model, the random slope model has two additional parameters: σ^2_{u1} and σ_{u01}. We can compare the random slope and random intercept models using a likelihood (LR) ratio test of the null hypothesis $H_0: \sigma^2_{u1} = \sigma_{u01} = 0$. The LR test statistic is dimply the drop in the deviance as we move from the simpler random-intercept model to the model complex random-slope model. This is calculated as

LR = 139239.06 - 138580.51 = 658.55

on two degrees of freedom. This is far above the 5% critical threshold of 5.99 and as such provides overwhelming evidence that the random slope model provides a better fit to the data than the random intercept model.

For illustration we plot the fitted trajectories for the first 20 individuals, after calculating the predicted values of **phf** (stored in **predrs**).

- From the **Model** menu, select **Predictions**
- Ensure that each of $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, $\boldsymbol{u_{0j}}$ and $\boldsymbol{u_{1j}}$ are checked at the bottom of the window
- In the **output from prediction to** drop-down box select **c12**

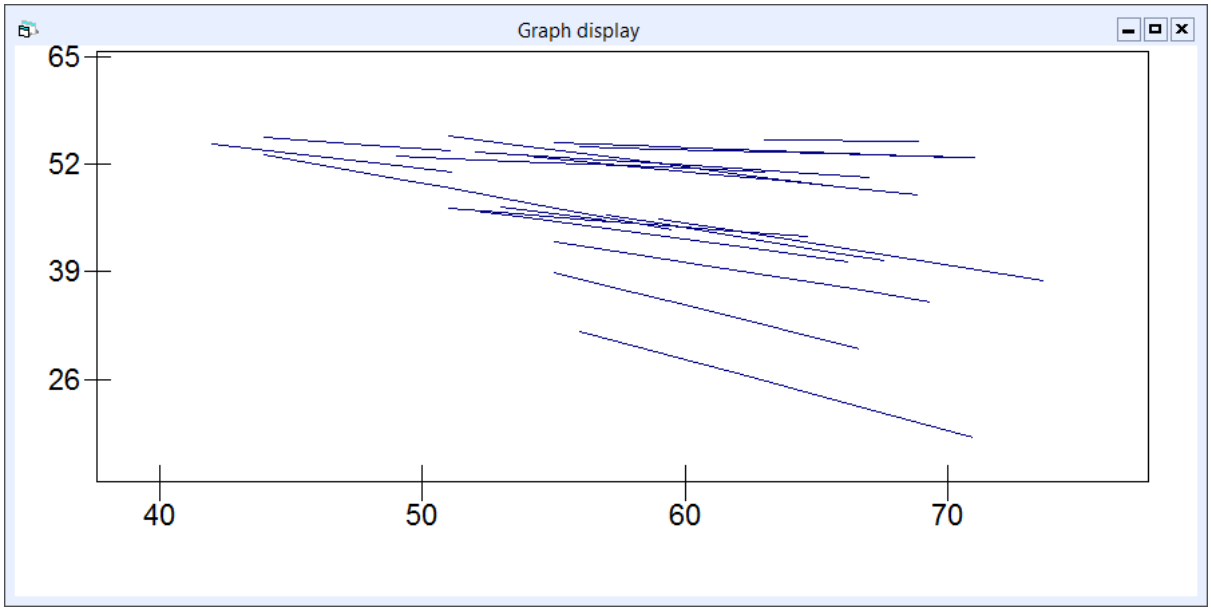
[4] It is of course possible that an individual's slope $\beta_1 + u_{1j}$ could be positive for sufficiently large u_{1j}.

- Press **Ctrl+N** to bring up the **Rename** window, enter **predrs** and click **Rename**
- Click **Calc**

Now that we have the predictions from the random slope model, we need to create an indicator variable to select the first 20 individuals and ignore the remaining 4407 individuals (id's 21 through 4427):

- From the **Data Manipulation** menu, select **recode** then select **by Range**
- In **Input columns** select **id** and in **Output columns** select **c13**
- Enter **1** into the first **Values in range** box, **20** into the second **Values in range** box and **1** into the **to new value** box then click **Add to action list**
- Enter **21** into the first **Values in range** box, **4427** into the second **Values in range** box and **0** into the **to new value** box then click **Add to action list**
- Click **Execute**
- Now that we have our new indicator variable created in **c13**, we can plot our graph
- From the **Graphs** menu, select **Customised Graph(s)**
- In the drop-down box at the top left of the window, select **D2** to specify a new display (MLwiN offers 10 potential displays in total, D1 through to D10)
- In the **plot what?** tab select **predrs** as the **y** variable and **age** as the **x** variable, select **c13** as the filter variable, **id** as the group variable, change **plot type** to **line**, and click **apply**. Note that it is by filtering on c13 that we restrict the plot to the first 20 individuals in the data.

After a little wait you should see the following graph.



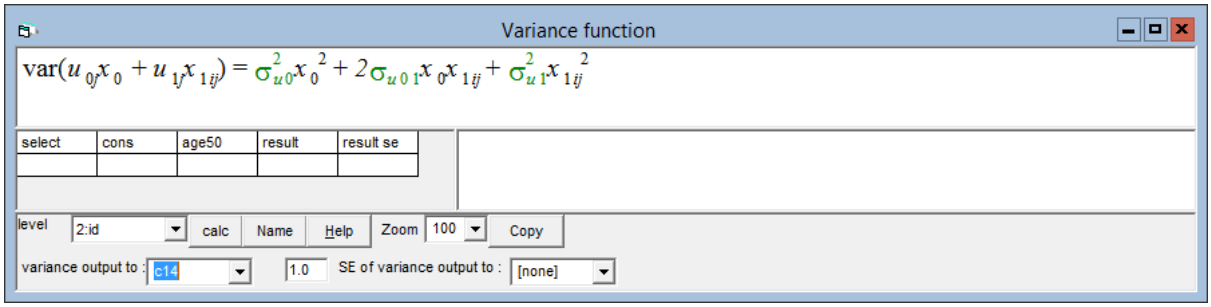
The predicted lines show a 'fanning out' pattern, which is consistent with the positive covariance between slopes and intercepts combined with the negative overall slope of age. This implies that the between-individual variance increases with age.

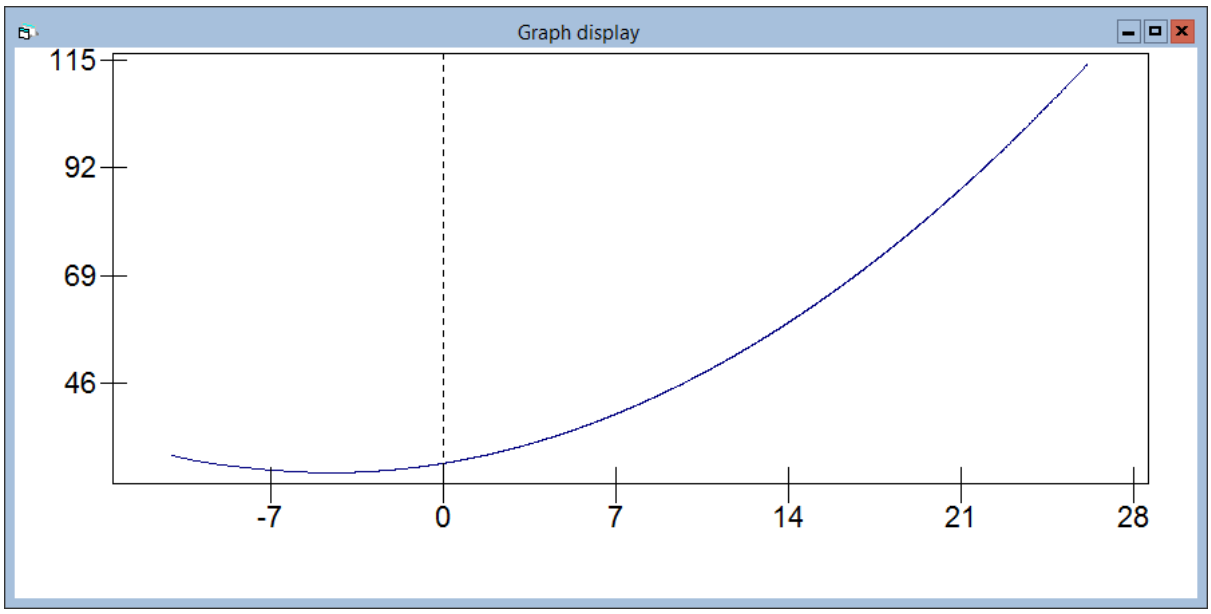
From equation (15.4) in C15.3.3 the between-individual variance is estimated as

$$\hat{\sigma}^2_{u0} + 2\,\hat{\sigma}_{u01}\text{age50}_{ij} + \hat{\sigma}^2_{u1}\text{age50}^2_{ij} = 28.82 + 0.86\,\text{age50}_{ij} + 0.09\,\text{age50}^2_{ij}$$

We can plot this function for values of **age50** between -10 and 25 (i.e. ages between 40 and 75) using the **Variance function** window.

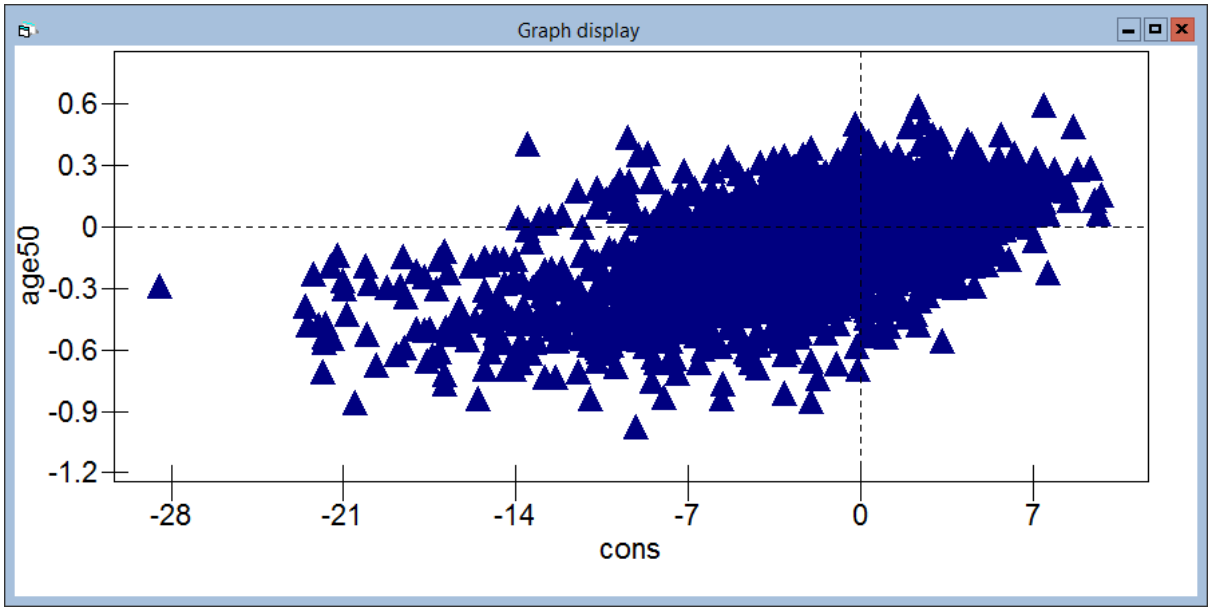
- From the **Model** menu, select **Variance function**
- In the **level** drop-down box at the bottom of the window, select **2:id** so that MLwiN knows to calculate the level 2 variance function
- In the **variance output to** drop-down box at the bottom of the window, select **c14**
- Check that your window matches that shown below, then click **calc**
- From the **Graphs** menu, select **Customised Graph(s)**
- In the drop-down box at the top left of the window, select **D3** to use another new display
- In the **plot what?** tab select **c14** as the **y** variable and **age50** as the **x** variable, change **plot type** to **line**, and click **apply**





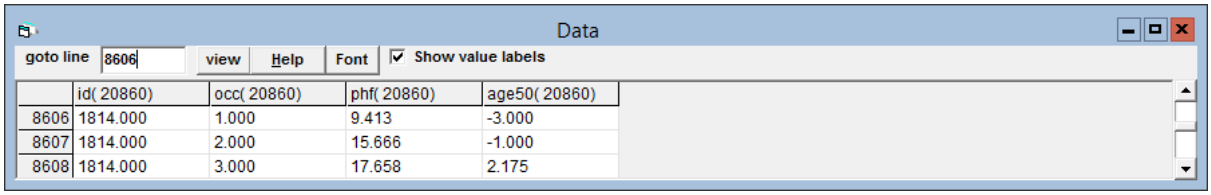
We next plot estimates of individuals slopes ($\hat{u}_{1j}$) versus individual intercepts ($\hat{u}_{0j}$). The random effect estimates are computed using the **Residuals** window in MLwiN.

- From the **Model** menu, select **Residuals**
- In the **level** drop-down box at the bottom select **2:id** so that the windows predicts the random intercept and slope effects at level-2 as opposed to the residuals at level-1.
- Click **Calc**
- Now click on the **Plots tab** and check the **residuals** button under the pairwise heading, then click **Apply**



As expected, given the positive estimate of σ_{u01}, there is a positive association between $\hat{u}_{1j}$ and $\hat{u}_{0j}$. There are some individuals with extreme negative intercepts including one extreme outlier. Clicking on the most extreme residual reveals that it is for individual 1814. Let us now examine the **phf** values for individual 1814.

- In the **Names** window, select **id**, **occ**, **phf** and **age50** and click **View**
- In the **goto line** box, enter **8606** to view individual 1814

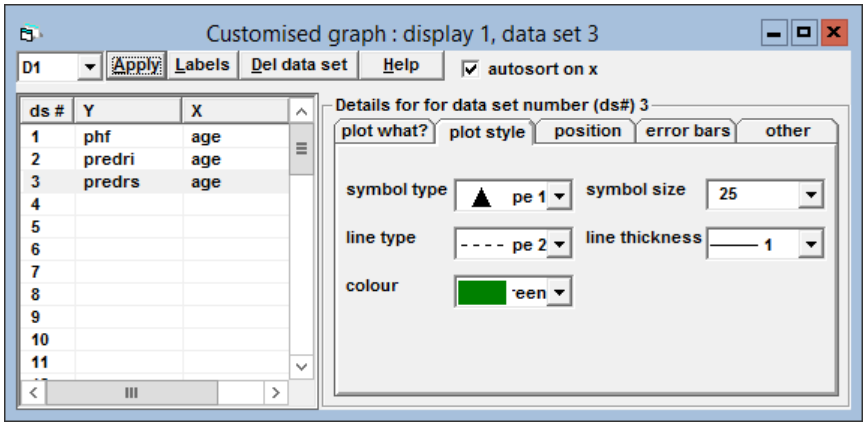


	id(20860)	occ(20860)	phf(20860)	age50(20860)
8606	1814.000	1.000	9.413	-3.000
8607	1814.000	2.000	15.666	-1.000
8608	1814.000	3.000	17.658	2.175

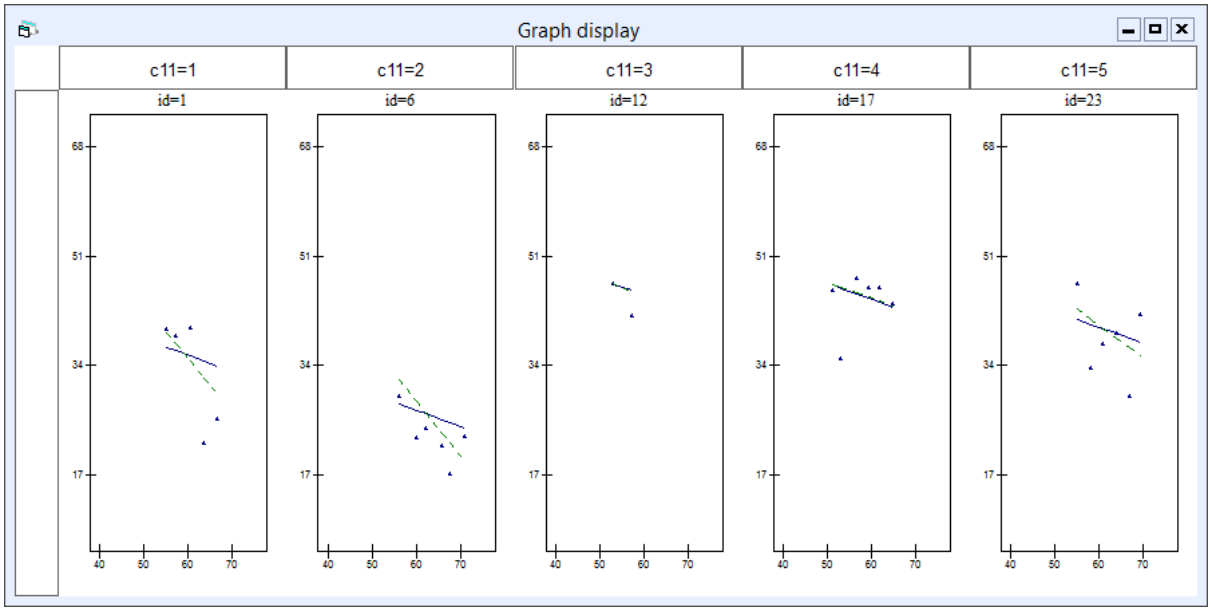
It can be seen that at each occasion this individual has physical functioning scores that are far below the mean (which was found to be close to 50 in P15.3.1).

To get an idea of the extent to which allowing random slopes improves predictions, fitted trajectories from the random intercept and random slope models are plotted for selected individuals. The fitted trajectories are superimposed on a scatterplot of the observed values of **phf** by **age**.

- From the **Graphs** menu, select **Customised Graph(s)**
- In the drop-down box at the top left of the window, select **D1** to return to the very first display
- Click on the third line of the left hand window (**ds#**) to add a new overlay to the existing graph
- In the **plot what?** tab select **predrs** as the **y** variable and **age** as the **x** variable, change **plot type** to **line**, and change **col codes** to **c11** (to again restrict the graph display to the first five individuals in the data)
- In the **plot style** tab change **line type** to **Type 2** and **colour** to **2 green**
- Check that the window matches that shown below, then click **apply**



After a short wait you should see the following graph:



Allowing slopes to vary across individuals brings the predicted lines closer to the observed values of **phf** for each individual.

Finally, we examine the within-individual correlations and standard deviations implied by the random slopes model for the first two individuals. We run the same series of commands as for the previous random-intercept model.

- From the **Data Manipulation** menu, select **Command interface**
- Click on **Output**
- Enter the following commands one block at a time and check that your output matches that shown below:

  VMAT 1 c100

  CALC g1 = c100

  PRIN g1
- **g1** in the output window shows the model-implied covariance matrix for individual 1 (**VMAT 1** tells MLwiN to use the first individual/cluster in the dataset)

  CALC g2 = SQRT(DIAG(g1))

  PRIN g2
- **g2** in the output window shows the model-implied standard deviations for individual 1

  CALC g3 = 1/g2

  CALC g4 = ~(g3*g1)*g3

  PRIN g4
- **g4** in the output window shows the model-implied correlation matrix for individual 1

  VMAT 2 c101

  CALC g5 = c101

  PRIN g5
- **g5** in the output window shows the model-implied covariance matrix for individual 2

  CALC g6 = SQRT(DIAG(g5))

  PRIN g6
- **g6** in the output window shows the model-implied standard deviations for individual 2

  CALC g7 = 1/g6

  CALC g8 = ~(g7*g5)*g7

  PRIN g8
- **g8** in the output window shows the model-implied correlation matrix for individual 2

The covariance matrix, vector of standard deviations and correlation matrix for the total residual for the first individual in the data are now:

```
->VMAT 1 c100
->CALC g1 = c100
->PRIN g1

      c1496         c1497         c1498         c1499         c1500
 N =      5             5             5             5             5
   1  63.943        37.183        40.222        43.098        45.742
   2  37.183        67.870        42.999        46.470        49.661
   3  40.222        42.999        76.282        52.230        56.356
   4  43.098        46.470        52.230        86.219        62.690
   5  45.742        49.661        56.356        62.690        97.053
```

```
->CALC g2 = SQRT(DIAG(g1))
->PRIN g2

      c1495
 N =      5
   1  7.9964
   2  8.2383
   3  8.7339
   4  9.2854
   5  9.8516
```

The corresponding covariance matrix, vector of standard deviations and correlation matrix for the second individual in the data are now:

```
->CALC g3 = 1/g2
->CALC g4 = ~(g3*g1)*g3
->PRIN g4

      c1489         c1490         c1491         c1492         c1493
 N =      5             5             5             5             5
   1  1.0000        0.56442       0.57591       0.58044       0.58064
   2  0.56442       1.0000        0.59760       0.60748       0.61189
   3  0.57591       0.59760       1.0000        0.64403       0.65497
   4  0.58044       0.60748       0.64403       1.0000        0.68532
   5  0.58064       0.61189       0.65497       0.68532       1.0000
```

```
->VMAT 2 c101
->CALC g5 = c101
->PRIN g5

      c1483         c1484         c1485         c1486         c1487         c1488
 N =      6             6             6             6             6             6
   1  65.814        41.202        43.296        46.954        48.629        51.955
   2  41.202        75.142        49.482        54.513        56.815        61.390
   3  43.296        49.482        81.320        58.543        61.181        66.421
   4  46.954        54.513        58.543        94.125        68.808        75.211
   5  48.629        56.815        61.181        68.808        100.84        79.233
   6  51.955        61.390        66.421        75.211        79.233        115.77
```

```
->CALC g6 = SQRT(DIAG(g5))
->PRIN g6

      c1482
 N =      6
   1  8.1126
   2  8.6684
   3  9.0177
   4  9.7018
   5  10.042
   6  10.759
```

```
->CALC g7 = 1/g6
->CALC g8 = ~(g7*g5)*g7
->PRIN g8

      c1475         c1476         c1477         c1478         c1479         c1480
 N =      6             6             6             6             6             6
   1  1.0000        0.58589       0.59182       0.59657       0.59693       0.59523
   2  0.58589       1.0000        0.63300       0.64820       0.65270       0.65822
   3  0.59182       0.63300       1.0000        0.66916       0.67562       0.68457
   4  0.59657       0.64820       0.66916       1.0000        0.70627       0.72051
   5  0.59693       0.65270       0.67562       0.70627       1.0000        0.73334
   6  0.59523       0.65822       0.68457       0.72051       0.73334       1.0000
```

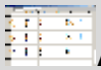
While the standard deviations were constant across occasions and individuals for the random intercepts model, the standard deviations now increase with age. As the within-individual variance is constant, this is due to the increase in the between-individual variance with age. We also see that the within-individual correlations depend on the age at each occasion, with higher correlations estimated for measurements at older ages. In the case of individual 2, for example, the correlation between occasions 5 and 6 (when the individual was age 67.6 and 71.0) is estimated as 0.73 compared to the correlation of 0.59 between occasions 1 and 2 (at ages 56 and 60). This could suggest greater fluctuations in physical functioning at younger ages, but we would need to examine trajectories for all individuals to explore this further.

P15.4 Nonlinear Growth

P15.4.1 Quadratic and higher-order polynomials

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to  ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "  15.4.1.wsz"

We have thus far assumed that physical health functioning declines linearly with age, but the plot of selected individual trajectories in P15.3.1 suggests a highly nonlinear relationship. In this section, we consider a quadratic model where age-squared is included as an additional predictor variable.

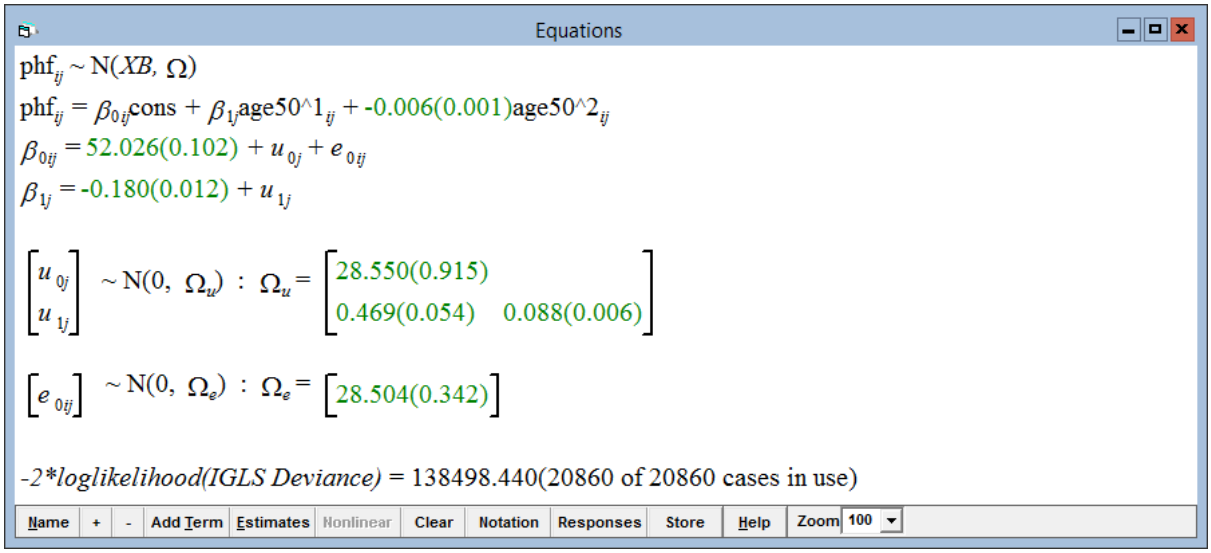
We begin by fitting a model with a fixed coefficient for age squared.

- From the **Model** menu, select **Equations** to bring up the **Equations** window with our previous model already specified
- Click on **age50** and click **Delete Term** to remove the previous centred age covariate from the model entirely so that we can go on to add centred age

again, but this time as a second-order polynomial (i.e., linear and quadratic terms)

- Click **Add Term**, select **age50** in the variable drop-down box, click the **polynomial** check box, select **2** in the **poly degree** drop-down box, and click **done** to add centred age as a the second-order polynomial
- Click on **age50^1** (the linear component) and check the **j(id)** box to re-specify the random slope for the **age50** term and then click **done**
- Click **Start** to run the model

You should obtain the following results:



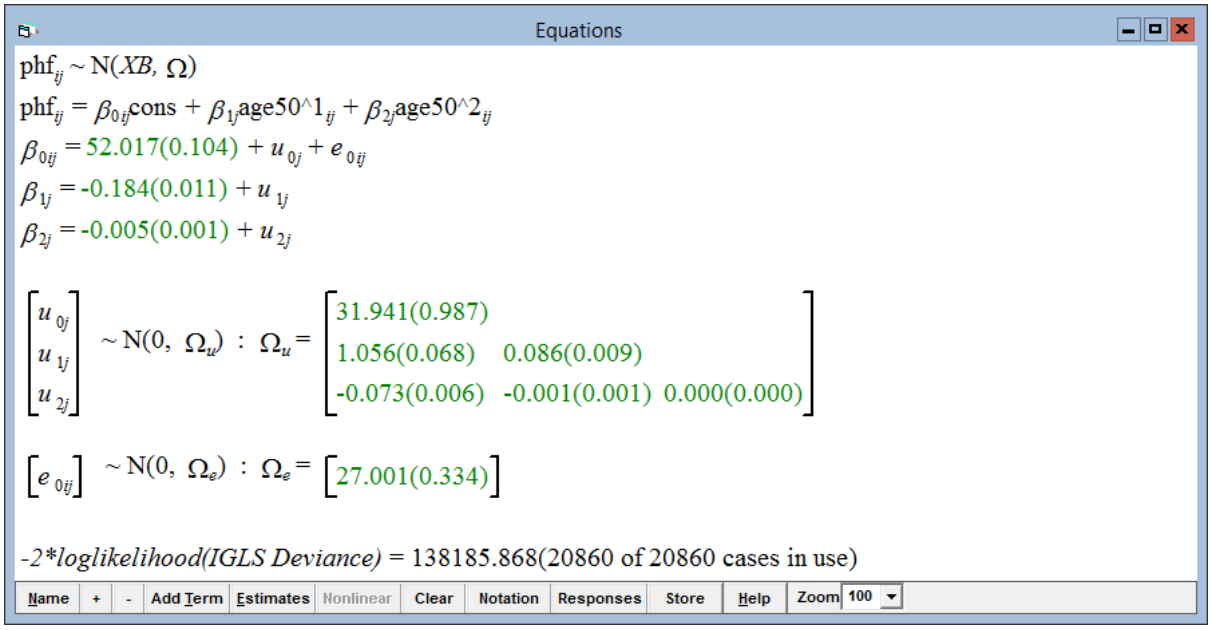
The negative coefficients for both **age50** and **age50^2**, and the significance of both coefficients, suggest that the decline in physical health functioning accelerates at older ages. From C15.4.1, the expected rate of change is given by

$$\frac{d\ \mathrm{phf}_{ij}}{d\ \mathrm{age50}_{ij}} = -0.180 - 2(0.006)\mathrm{age50}_{ij}$$

Thus, for example, a 1-year increase in age is associated with decrease of 0.18 points for a 50 year old and 0.42 points for a 70 year old (**age50** = 20).

We now extend the model to allow the coefficient of age-squared to also vary across individuals:

- In the **Equations** window click on the **age50^2** term, check the **j(id)** check box to allow the age squared term to vary across individuals and click **done**
- Click **Start** to run the model



The above model takes the form of equation (15.8) in C15.4.1:

$$\begin{aligned}
\text{phf}_{ij} &= \beta_{0j} + \beta_{1j}\text{age50}_{ij} + +\beta_{2j}\text{age50}^2_{ij} + e_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j} \\
\beta_{2j} &= \beta_2 + u_{2j}
\end{aligned}$$

$$e_{ij} \sim N(0, \sigma^2_e) \text{ and } \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N(\mathbf{0}, \Omega_u) \text{ where } \Omega_u = \begin{pmatrix} \sigma^2_{u0} & & \\ \sigma_{u01} & \sigma^2_{u1} & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} \end{pmatrix}$$

Compared to the previous model, there are three new random effects parameters, all given in the bottom row of Ω_u. We test whether they are simultaneously equal to zero using a likelihood ratio test. The LR test statistic is calculated as the reduction in deviance as we move from the simpler previous model to the more complex current model.

LR = 138498.44 - 138185.87 = 312.57

The test statistic of 138498 on three degrees of freedom is far above the 5% critical threshold of 7.82 so the null hypothesis that all three new parameters equal zero is strongly rejected, and the full quadratic model is preferred.

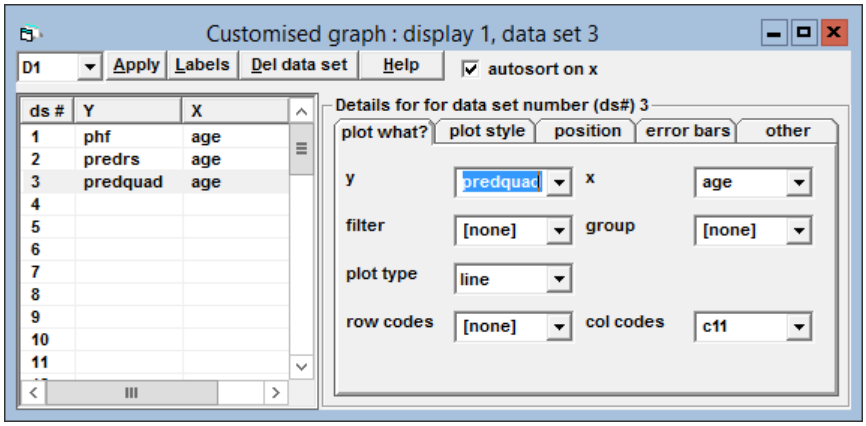
To get an idea of how adding a quadratic term (with random effect) improves model fit, we plot the predicted trajectories from the quadratic and linear models alongside the observed values of **phf** for the same 5 individuals as before. The data file includes the variable **predrs** in which the predictions from the linear random slopes model are stored. To calculate predictions from our current model:

- From the **Model** menu, select **Predictions**

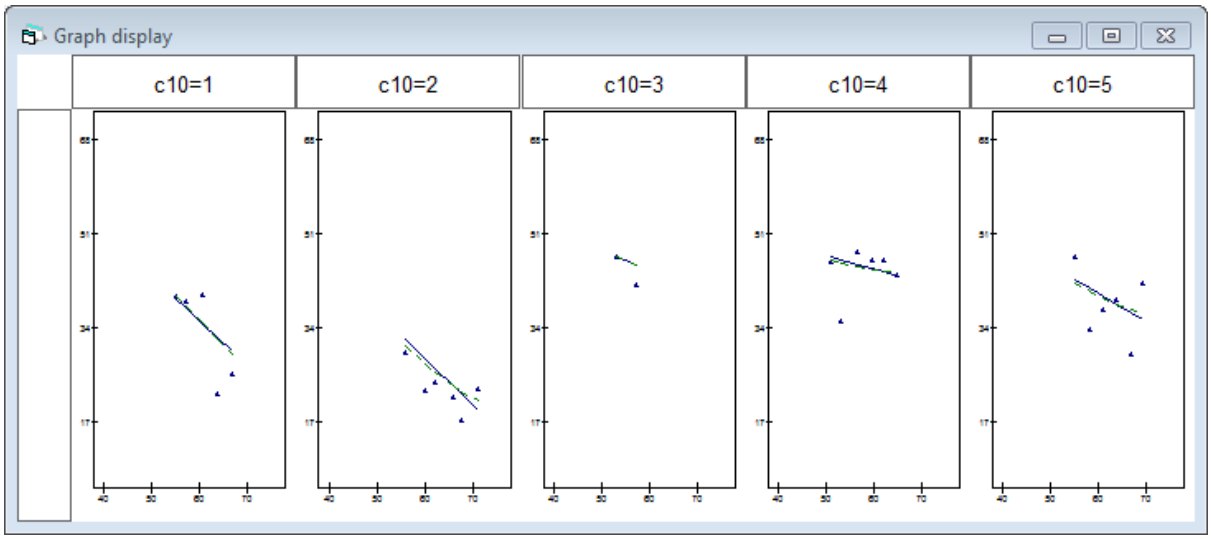
- Ensure that every term at the bottom of the window is checked, with the exception of the residuals e_{0ij}
- In the **output from prediction to** drop-down box select **c13**
- Press **Ctrl+N** to bring up the **Rename** window, enter **predquad** and click **Rename**
- Click **Calc**

We will now plot the predictions from the current quadratic model alongside the predictions from the previous linear model from P15.3.3 (this is already saved in the worksheet):

- From the **Graphs** menu, select **Customised Graph(s)**
- Click on the second line of the left hand window (**ds#**) and change the **y** variable to **predrs**
- Click on the third line of the left hand window (**ds#**) and change the **y** variable to **predquad**
- Check that the window matches that shown below, then click **Apply**



After a short wait you should see the following graph:



Although there is some suggestion that the prediction lines from the quadratic model lie closer to the data points than the linear predictions, the improvement in fit appears small for these individuals. The highly nonlinear trajectories would require a higher-order polynomial. Fitting a cubic age term, with associated random effect, is left as an exercise for the reader. You should find that this model converges, but that standard errors could not be computed for the random effect parameters. This is a fairly common occurrence for models with multiple random effects, and suggests that the model should be simplified.

We end by inspecting the implied correlation matrix for the first two individuals. We calculate these matrices using the same series of commands as for the previous two models.

- From the **Data Manipulation** menu, select **Command interface**
- Click on **Output**
- Enter the following commands one block at a time and check that your output matches that shown below:

  VMAT 1 c100

  CALC g1 = c100

  PRIN g1
- **g1** in the output window shows the model-implied covariance matrix for individual 1 (**VMAT 1** tells MLwiN to use the first individual/cluster in the dataset)

  CALC g2 = SQRT(DIAG(g1))

  PRIN g2
- **g2** in the output window shows the model-implied standard deviations for individual 1

  CALC g3 = 1/g2

  CALC g4 = ~(g3*g1)*g3

  PRIN g4
- **g4** in the output window shows the model-implied correlation matrix for individual 1

  VMAT 2 c101

  CALC g5 = c101

  PRIN g5
- **g5** in the output window shows the model-implied covariance matrix for individual 2

  CALC g6 = SQRT(DIAG(g5))

  PRIN g6
- **g6** in the output window shows the model-implied standard deviations for individual 2

  CALC g7 = 1/g6

  CALC g8 = ~(g7*g5)*g7

  PRIN g8

- **g8** in the output window shows the model-implied correlation matrix for individual 2

```
->VMAT 1 c100
->CALC g1 = c100
->PRIN g1

      c1496         c1497         c1498         c1499         c1500
 N =     5             5             5             5             5
   1  67.903        42.096        42.811        41.950        39.839
   2  42.096        70.686        45.160        45.118        43.844
   3  42.811        45.160        75.182        49.887        50.467
   4  41.950        45.118        49.887        80.657        56.481
   5  39.839        43.844        50.467        56.481        88.795
```

```
->CALC g2 = SQRT(DIAG(g1))
->PRIN g2

      c1495
 N =     5
   1  8.2404
   2  8.4075
   3  8.6708
   4  8.9809
   5  9.4231
```

```
->CALC g3 = 1/g2
->CALC g4 = ~(g3*g1)*g3
->PRIN g4

      c1489         c1490         c1491         c1492         c1493
 N =     5             5             5             5             5
   1  1.0000        0.60761       0.59917       0.56685       0.51307
   2  0.60761       1.0000        0.61948       0.59753       0.55342
   3  0.59917       0.61948       1.0000        0.64063       0.61767
   4  0.56685       0.59753       0.64063       1.0000        0.66740
   5  0.51307       0.55342       0.61767       0.66740       1.0000
```

```
->VMAT 2 c101
->CALC g5 = c101
->PRIN g5

      c1483         c1484         c1485         c1486         c1487         c1488
 N =     6             6             6             6             6             6
   1  69.335        43.994        43.970        42.410        41.051        37.152
   2  43.994        74.608        48.805        49.685        49.573        48.391
   3  43.970        48.805        77.805        53.328        54.072        54.788
   4  42.410        49.685        53.328        86.298        61.860        66.641
   5  41.051        49.573        54.072        61.860        92.394        72.352
   6  37.152        48.391        54.788        66.641        72.352        111.24
```

```
->CALC g6 = SQRT(DIAG(g5))
->PRIN g6

      c1482
 N =     6
   1  8.3267
   2  8.6376
   3  8.8207
   4  9.2897
   5  9.6122
   6  10.547
```

```
->CALC g7 = 1/g6
->CALC g8 = ~(g7*g5)*g7
->PRIN g8

      c1475         c1476         c1477         c1478         c1479         c1480
 N =      6             6             6             6             6             6
   1  1.0000        0.61168       0.59866       0.54826       0.51290       0.42305
   2  0.61168       1.0000        0.64057       0.61921       0.59708       0.53119
   3  0.59866       0.64057       1.0000        0.65080       0.63775       0.58892
   4  0.54826       0.61921       0.65080       1.0000        0.69277       0.68018
   5  0.51290       0.59708       0.63775       0.69277       1.0000        0.71369
   6  0.42305       0.53119       0.58892       0.68018       0.71369       1.0000
```

For ease of comparison, the table below shows the estimated correlations and standard deviations from the linear and quadratic models for individual 2. There are some large differences in both between the two models. One notable difference is that the estimated correlations from the quadratic model follow a more plausible pattern whereby the correlation between any pair of measurements decreases with the length of time between them. The estimated between-individual standard deviation increases with age for both models, but with a smaller increase between ages 62.1 and 65.9 years and a larger increase between ages 67.6 and 71 years for the quadratic model than for the linear model. Again, the pattern estimated by the quadratic model seems the most plausible: we might expect the between-individual variance to not only increase with age, but to increase more rapidly at older ages.

Within-individual correlations for individual 2 (top=linear, ***bottom=quadratic***)						
Occasion (age in years)	1	2	3	4	5	6
1 (56)	1					
2 (60)	0.586 ***0.612***	1				
3 (62.1)	0.592 ***0.599***	0.633 ***0.641***	1			
4 (65.9)	0.597 ***0.548***	0.648 ***0.619***	0.669 ***0.651***	1		
5 (67.6)	0.597 ***0.513***	0.653 ***0.597***	0.676 ***0.638***	0.706 ***0.693***	1	
6 (71.0)	0.595 ***0.423***	0.658 ***0.531***	0.685 ***0.589***	0.721 ***0.680***	0.733 ***0.714***	1
Within-individual standard deviations						
Occasion	1	2	3	4	5	6
SD	8.113 ***8.327***	8.668 ***8.638***	9.018 ***8.821***	9.702 ***9.290***	10.042 ***9.612***	10.760 ***10.547***

P15.4.2 Splines

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to  ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).

- Click "  15.4.2.wsz"

The quadratic model considered above assumes the same quadratic function across time (the observed age range in our physical health functioning example). Splines offer greater flexibility as they allow a different polynomial to be specified for different intervals of time. The boundaries of the time intervals, which must be chosen by the researcher, are called knots or join points.

In C15.4.2 a simple linear spline function was considered, and applied to simulated data with 3 time intervals (and therefore 2 knots). We now replicate this analysis. Recall that the dataset has 20 measurements per individual (coded $t_{ij} = 0, 1, \ldots, 19$) and a piecewise linear spline is fitted with knots at $t_1 = 6$ and $t_2 = 13$. There are three spline variables defined as

$$s_{1ij} = \begin{cases} t_{ij} & t_{ij} \leq 6 \\ 6 & t_{ij} > 6 \end{cases}$$

$$s_{2ij} = \begin{cases} 0 & t_{ij} \leq 6 \\ t_{ij} - 6 & 6 < t_{ij} \leq 13 \\ 7 & t_{ij} > 13 \end{cases}$$

$$s_{3ij} = \begin{cases} 0 & t_{ij} \leq 13 \\ t_{ij} - 13 & t_{ij} > 13 \end{cases}$$

These variables are included in the data file "15.4.2.wsz", called **s1**, **s2** and **s3**. The original time variable is **t**, and the individual identifier is **id**. We obtain summary statistics for all variables and view **id**, **t** and the spline variables for the first individual. The coding of **s1**, **s2** and **s3** is the same as that shown in Table 15.10 of C15.4.2.

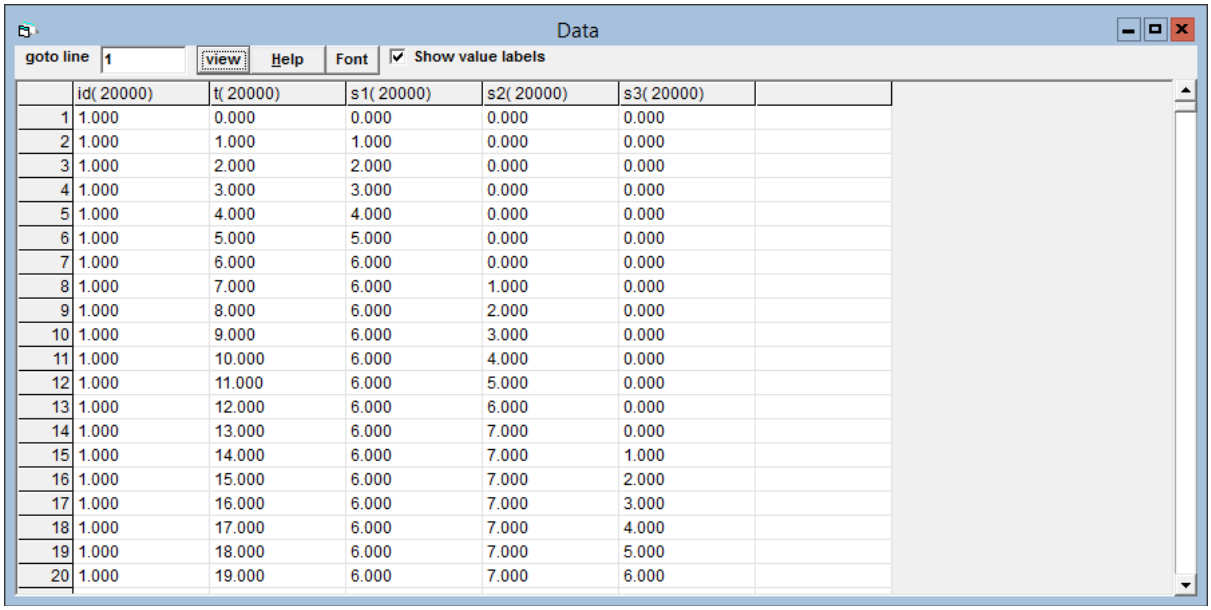
It is helpful to look at summary statistics for all variables in the dataset.

- From the **Basic Statistics** menu, select **Averages and Correlations**
- Highlight all of the variables except for **cons** and click **Calculate**

```
->AVERage 6  'id'   't'   's1'   's2'   's3'   'y'

                 N     Missing    Mean         s.d.
id     20000          0     500.50        288.68
t      20000          0     9.5000        5.7664
s1      20000         0      4.9500        1.8568
s2      20000         0      3.5000        3.0742
s3      20000         0      1.0500        1.8568
y      20000          0     6.6689        4.4665
```

Similarly, we can browse the data in the usual way.

- In the **Names** window, highlight **id**, **t**, **s1, s2** and **s3** then click **View**



Data

goto line 1 | view | Help | Font | Show value labels

	id(20000)	t(20000)	s1(20000)	s2(20000)	s3(20000)
1	1.000	0.000	0.000	0.000	0.000
2	1.000	1.000	1.000	0.000	0.000
3	1.000	2.000	2.000	0.000	0.000
4	1.000	3.000	3.000	0.000	0.000
5	1.000	4.000	4.000	0.000	0.000
6	1.000	5.000	5.000	0.000	0.000
7	1.000	6.000	6.000	0.000	0.000
8	1.000	7.000	6.000	1.000	0.000
9	1.000	8.000	6.000	2.000	0.000
10	1.000	9.000	6.000	3.000	0.000
11	1.000	10.000	6.000	4.000	0.000
12	1.000	11.000	6.000	5.000	0.000
13	1.000	12.000	6.000	6.000	0.000
14	1.000	13.000	6.000	7.000	0.000
15	1.000	14.000	6.000	7.000	1.000
16	1.000	15.000	6.000	7.000	2.000
17	1.000	16.000	6.000	7.000	3.000
18	1.000	17.000	6.000	7.000	4.000
19	1.000	18.000	6.000	7.000	5.000
20	1.000	19.000	6.000	7.000	6.000

The model takes the form

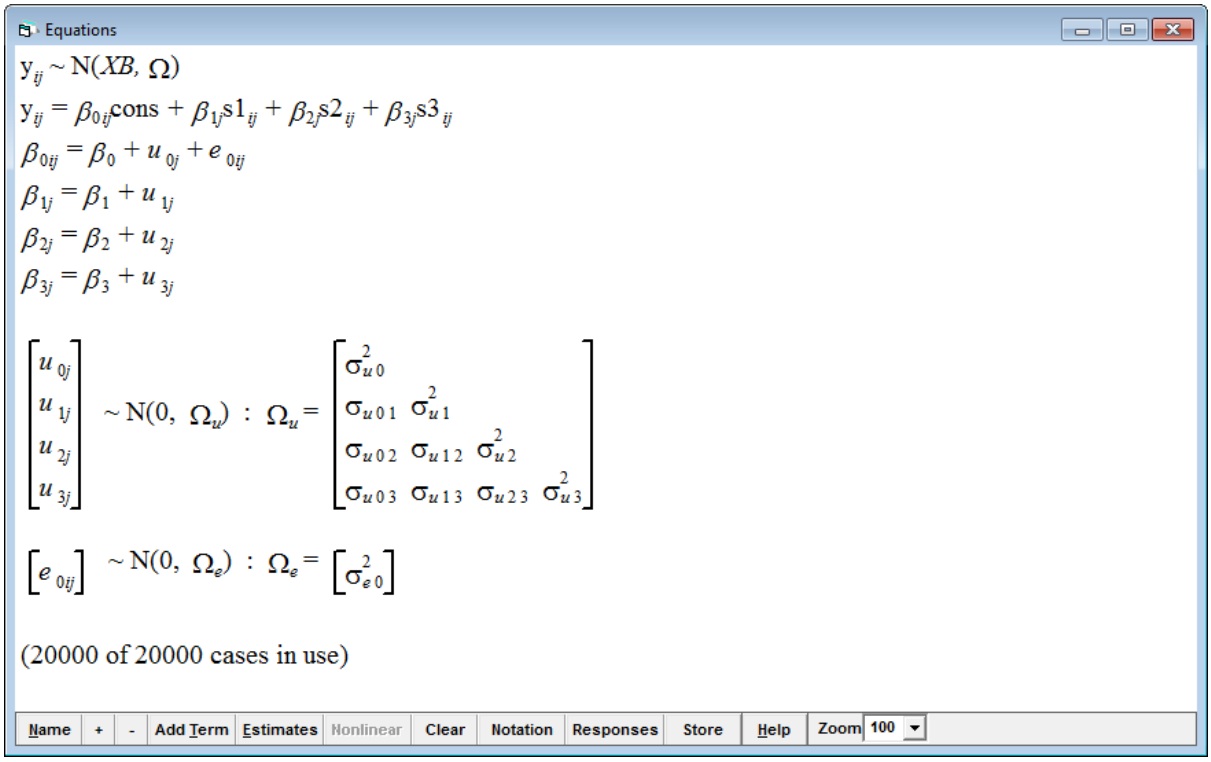
$$y_{ij} = \beta_{0j} + \beta_{1j}s_{1ij} + \beta_{2j}s_{2ij} + \beta_{3j}s_{3ij} + e_{ij}$$

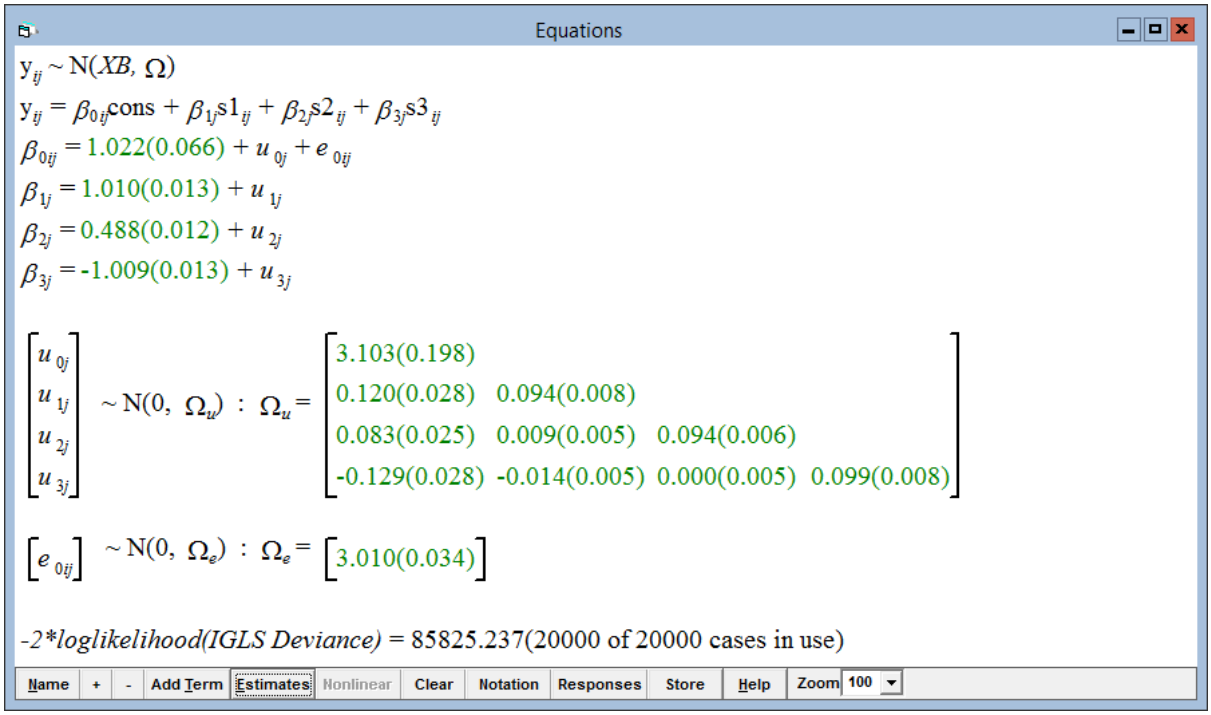
where $\beta_{kj} = \beta_k + u_{kj}$ $(k = 0, 1, \ldots, 3)$ and the u_{kj} follow a 4-dimensional multivariate normal distribution with covariance matrix

$$\Omega_u = \begin{pmatrix} \sigma^2_{u0} & & & \\ \sigma_{u01} & \sigma^2_{u1} & & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma^2_{u3} \end{pmatrix}$$

Next we shall specify and fit the model in the Equations window.

- From the **Model** menu, select **Equations**
- Click on the red **y**, select **y** in the **y** drop-down box, **2-ij** in the **N levels** drop-down box, **id** in **level 2(j)**, **t** in **level 1(i)**, then click **done**
- Click on the red $\boldsymbol{\beta_0}$ term, select **cons** from the drop-down box, check the **j(id)** and **i(id)** boxes, and click **Done**
- Click **Add Term** and select **s1** from the **variable** drop-down box
- Click **Add Term** and select **s2** from the **variable** drop-down box
- Click **Add Term** and select **s3** from the **variable** drop-down box
- Click on each of these terms and check the **j(id)** check boxes to allow their slope coefficients to vary random across individuals, then click on the **+** button twice to show the full model (this should match the first screenshot below)
- Click **Start** to run the model, then click the **Estimates** button twice after the model has converged to reveal the parameter estimates and their standard errors



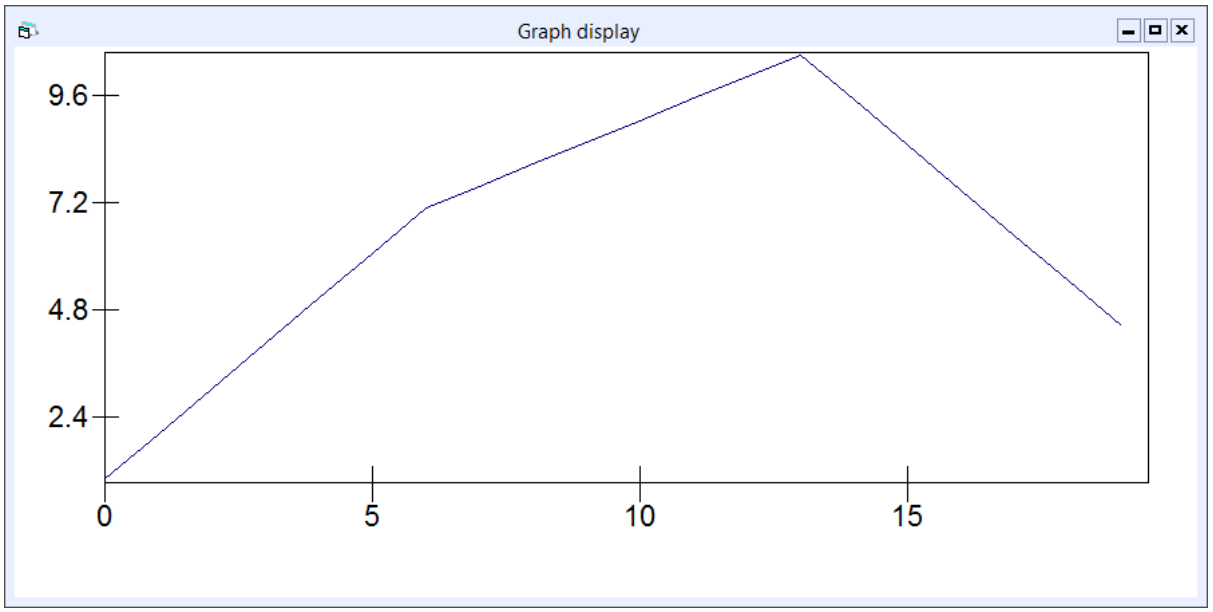


The results agree with those in Table 15.12 of C15.4.2. To obtain a plot of the mean predicted value of y at each value of time, we first obtain the predictions using MLwiN's **Predictions** window before plotting these in a graph.

- From the **Model** menu, select **Predictions**
- Click on **fixed** and select **Include all fixed coefficients** to base the prediction on only the fixed-part of the model
- In the **output from prediction to** drop-down box select **c8**

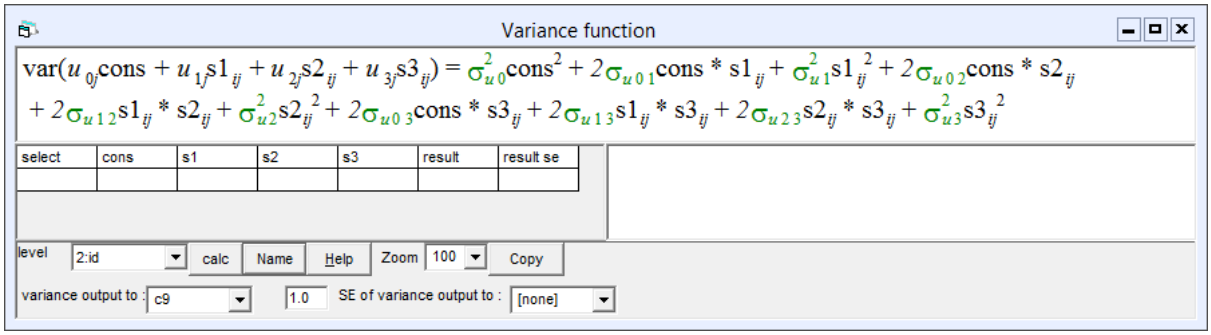
- Press **Ctrl+N** to bring up the **Rename** window, enter **predy** and click **Rename**
- Click **Calc**
- From the **Graphs** menu, select **Customised Graph(s)**
- In the **plot what?** tab select **predy** as the **y** variable and **t** as the **x** variable, and change **plot type** to **line**
- Click **Apply**

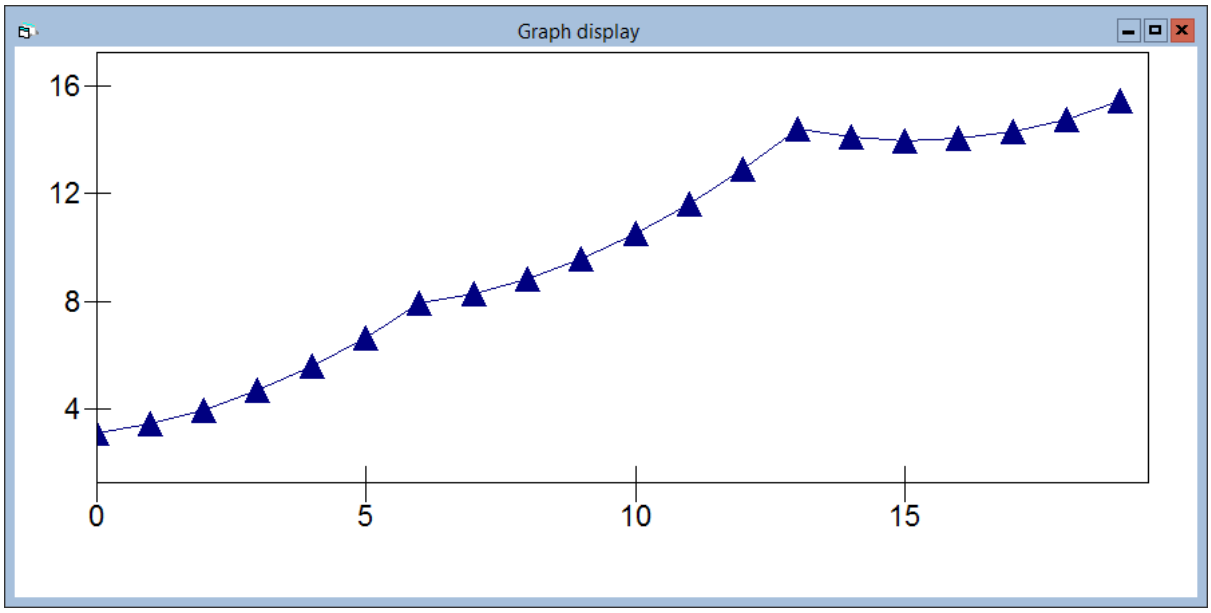
The resulting Graph display shows the predicted mean value of y against time.



We will now plot the between-individual variance at each occasion using the **Variance function** window.

- From the **Model** menu, select **Variance function**
- In the **level** drop-down box at the bottom of the window, select **2:id** so that the resulting variance function is calculated at the individual level as opposed to at level-1, the occasion level
- In the **variance output to** drop-down box at the bottom of the window, select **c9**
- Click **calc**
- From the **Graphs** menu, select **Customised Graph(s)**
- In the **plot what?** tab change the **y** variable to **c9**, change **plot type** to **line+point**,
- Check that the window matches that shown below, then click **Apply**





The variance increases with occasion, apart from some fluctuations after occasion 14 (the final interval of the spline).

Should we wish to calculate the estimated standard deviations of the total residual we must calculate the square root of the sum of the between-individual and within-individual variance at each occasion. For example, for $t_{ij} \leq 6$ (referred to as interval 1 in C15.4.2) the expression for the between-individual variance is

$$\sigma^2_{u0} + 2\sigma_{u01}t_{ij} + \sigma^2_{u1}t^2_{ij}$$

which is estimated as

$$3.103 + 0.240\ t_{ij} + 0.094\ t^2_{ij}$$

and the within-individual variance is estimated as 3.010.

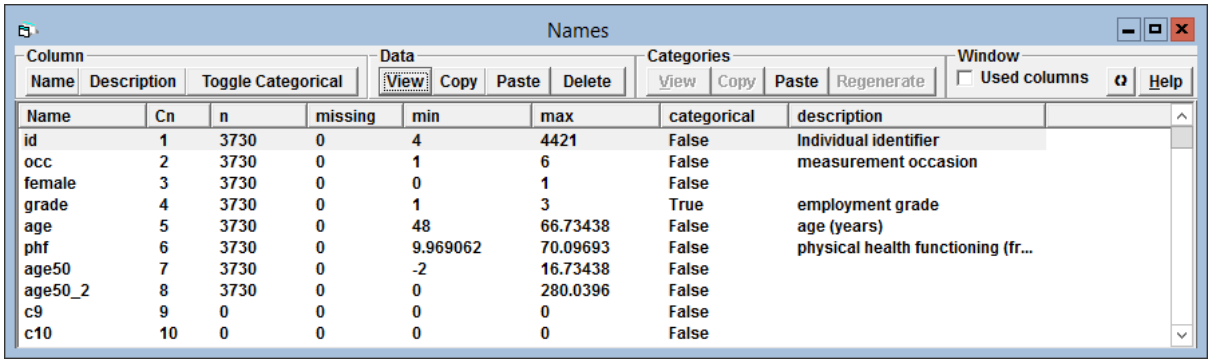
Thus the total variance at $t_{ij} = 0$ is 3.103 + 3.010 = 6.113, which gives a standard deviation of 2.47.

P15.4.3 Treating time as categorical: Multivariate response models

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "15.4.3.wsz"



Name	Cn	n	missing	min	max	categorical	description
id	1	3730	0	4	4421	False	Individual identifier
occ	2	3730	0	1	6	False	measurement occasion
female	3	3730	0	0	1	False	
grade	4	3730	0	1	3	True	employment grade
age	5	3730	0	48	66.73438	False	age (years)
phf	6	3730	0	9.969062	70.09693	False	physical health functioning (fr...
age50	7	3730	0	-2	16.73438	False	
age50_2	8	3730	0	0	280.0396	False	
c9	9	0	0	0	0	False	
c10	10	0	0	0	0	False	

The final nonlinear growth model we consider treats an individual's repeated measurements of y as a multivariate response. It is the most flexible model because it has parameters for the mean and variance of y at each occasion, and for the covariances between each pair of measurements. However, as discussed in C15.4.3, the multivariate model is only appropriate when measurement occasions are fixed, i.e. $t_{ij} = t_i$ for all individuals. The model is not suitable for the analysis of physical health functioning where functioning depends strongly on age and there is substantial variation between individuals' ages at each measurement occasions. In such observations, age and occasion are not interchangeable and age is the more appropriate time metric.

In order to illustrate the multivariate model, an approximate birth cohort has been constructed from the Whitehall II study by selecting individuals between ages 48 and 51 years (inclusive) at occasion 1.[5] To obtain summary statistics for age at each occasion for this subsample, we use the **Tabulate** window.

- From the **Basic Statistics** menu, select **Tabulate**
- Click the **Means** button
- In the **Variate column** drop-down box select **age**
- In the **Columns** drop-down box select **occ** so that the mean and standard deviation of age is reported separately by occasion

[5] Recall that the mean age at occasion 1 for all respondents is 50 years. We take an interval around the mean to increase the number of respondents in the analysis sample.

```
->TABUlate  'age'    'occ'

Variable tabulated is age

              1          2          3          4          5          6     TOTALS

      N     727        664        602        577        579        581       3730
  MEANS  49.407     52.449     55.733     58.708     61.150     64.059     56.513
   SD'S   1.083      1.255      1.229      1.211      1.226      1.203      5.223
```

We can see that the variation in age increases after occasion 1, but this is ignored when using occasion as the time metric in the multivariate model. In treating occasions as fixed we assume that variations in age at a given occasion and differences between individuals in the age gap between occasions are unimportant. These assumptions are made purely for purposes of illustration. In practice, the multivariate model would not be recommended for this study design.

Before fitting the multivariate model we obtain the mean of **phf** by occasion and the within-individual covariance matrix. As the multivariate model includes a parameter for each mean, variance and covariance, the model should approximately reproduce these estimates.[6] It is easier to carry out these computations when the data are in wide form, so we first remove unwanted long form data and then use the **Unsplit records** window to restructure the data.

- From the **Data Manipulation menu, select Command Interface** and run the following commands

  ERAS c3 c4 c5 c7 c8

  MOVE

  to first drop all variables other than the individual and occasion identifies and the phf scores themselves (currently stored in column **c6**), and to then move the phf variable itself to the third column in the worksheet
- From the **Data Manipulation** menu, select **Unsplit Records**
- In the **Occasion ID** drop-down box select the long form **occ** variable
- In the **Input case ID** drop-down box select the long form **id** variable
- In the **Output case ID** drop-down box select the empty column **c4** which will in turn be given the wide form version of the **id** variable
- In the **Occasion 1… Occasion 6** boxes select the empty columns **c5… c10** which in turn will store the values of **phf** in wide form
- In the **Unstacked from** drop-down box select the long form **phf variable**
- Check that the window matches that shown below, then click **Unstack**
- From the **Data Manipulation menu, select Command Interface** and run the following commands to rename the new wide form variables.

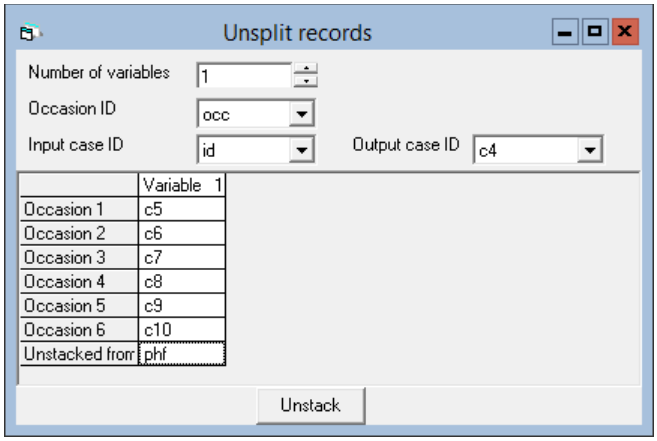
  NAME c4 'idwide'

  NAME c5 'phf1'

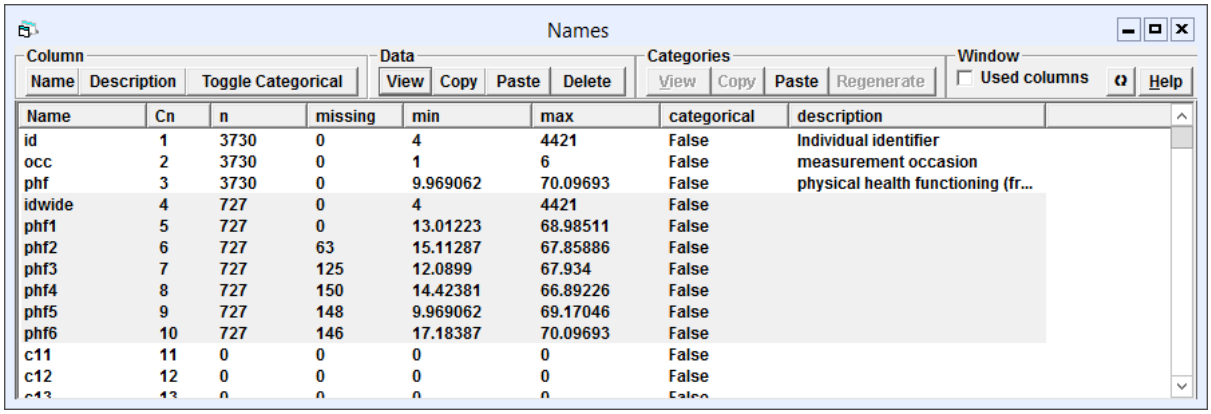
  NAME c6 'phf2'

[6] The parameter estimates from the model will be exactly equal to the descriptive statistics when there is no missing data, i.e. all individuals are present at each occasion.

NAME c7 'phf3'
NAME c8 'phf4'
NAME c9 'phf5'
NAME c10 'phf6'



The **Names** window should now look like this:



Name	Cn	n	missing	min	max	categorical	description
id	1	3730	0	4	4421	False	Individual identifier
occ	2	3730	0	1	6	False	measurement occasion
phf	3	3730	0	9.969062	70.09693	False	physical health functioning (fr...
idwide	4	727	0	4	4421	False	
phf1	5	727	0	13.01223	68.98511	False	
phf2	6	727	63	15.11287	67.85886	False	
phf3	7	727	125	12.0899	67.934	False	
phf4	8	727	150	14.42381	66.89226	False	
phf5	9	727	148	9.969062	69.17046	False	
phf6	10	727	146	17.18387	70.09693	False	
c11	11	0	0	0	0	False	
c12	12	0	0	0	0	False	

We can now calculate the means and standard deviations of the occasion-specific **phf** variables and the pairwise correlations between occasions.[7]

- From the **Basic Statistics menu, select Averages and Correlations**
- Click on the **Correlation** button
- Select the variables **phf1**-**phf6**
- Click **Calculate**

[7] Note that these correlations are based on the subsample of individuals with no missing values (i.e., the listwise or casewise deleted sample). This contrasts the Stata version of this practical where we reported pairwise correlations based on all available data.

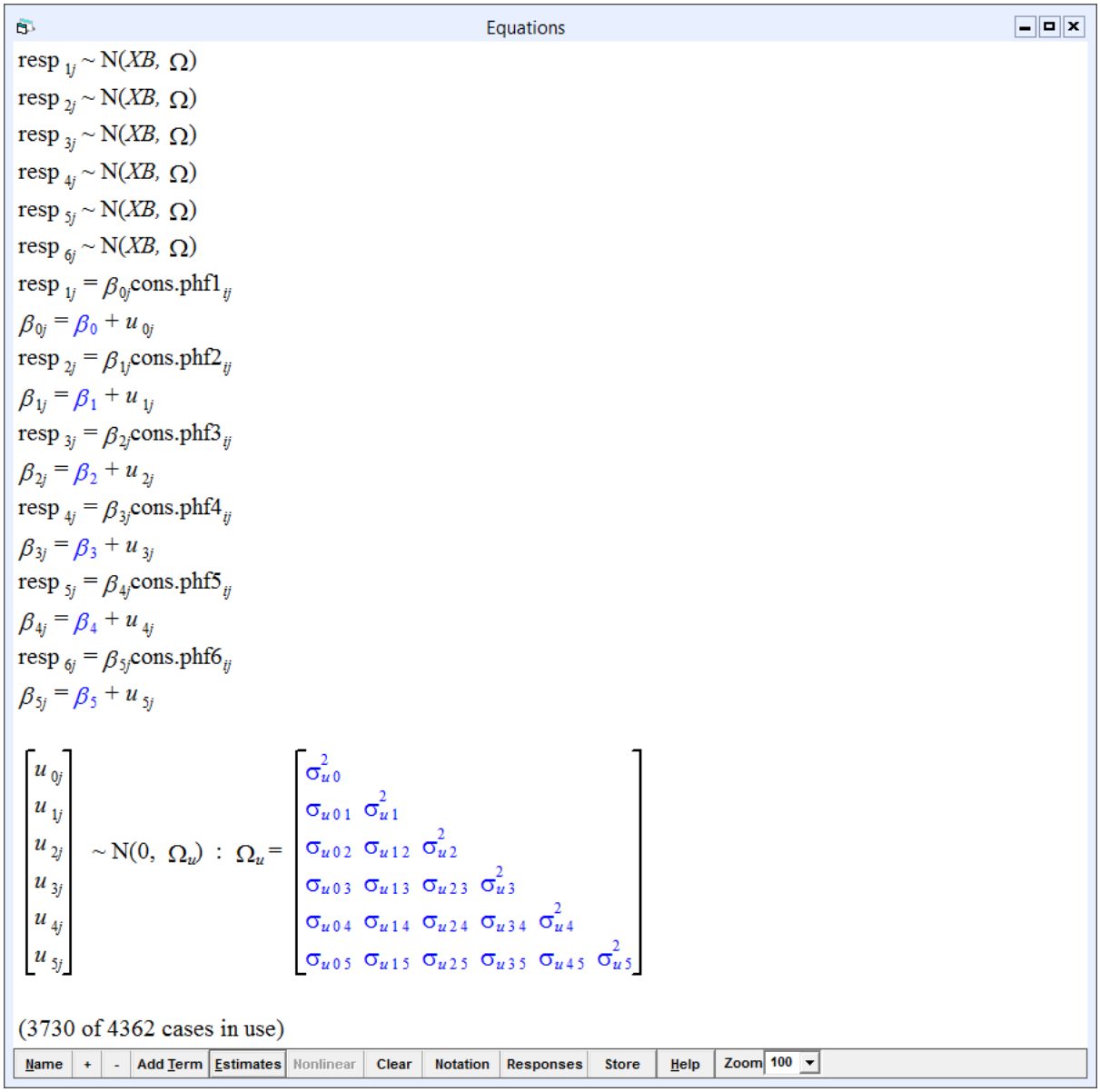
```
->CORMatrix 6  'phf1'   'phf2'   'phf3'   'phf4'   'phf5'   'phf6'

311 missing value(s)

416 observations

Means

phf1         phf2         phf3         phf4         phf5         phf6
52.876       51.366       51.471       51.396       49.542       49.745

S.D.'s

phf1         phf2         phf3         phf4         phf5         phf6
6.9188       7.8295       7.6289       7.3084       8.4269       8.3633

Correlations

              phf1         phf2         phf3         phf4         phf5         phf6

phf1          1.0000
phf2          0.5629       1.0000
phf3          0.6255       0.5902       1.0000
phf4          0.6120       0.5053       0.6227       1.0000
phf5          0.5057       0.4747       0.6365       0.5864       1.0000
phf6          0.4673       0.4243       0.5514       0.5253       0.6251       1.0000
```

We are now ready to fit the multivariate model, but we must once again create a variable named **cons** as a series of 1's for the model constant term:

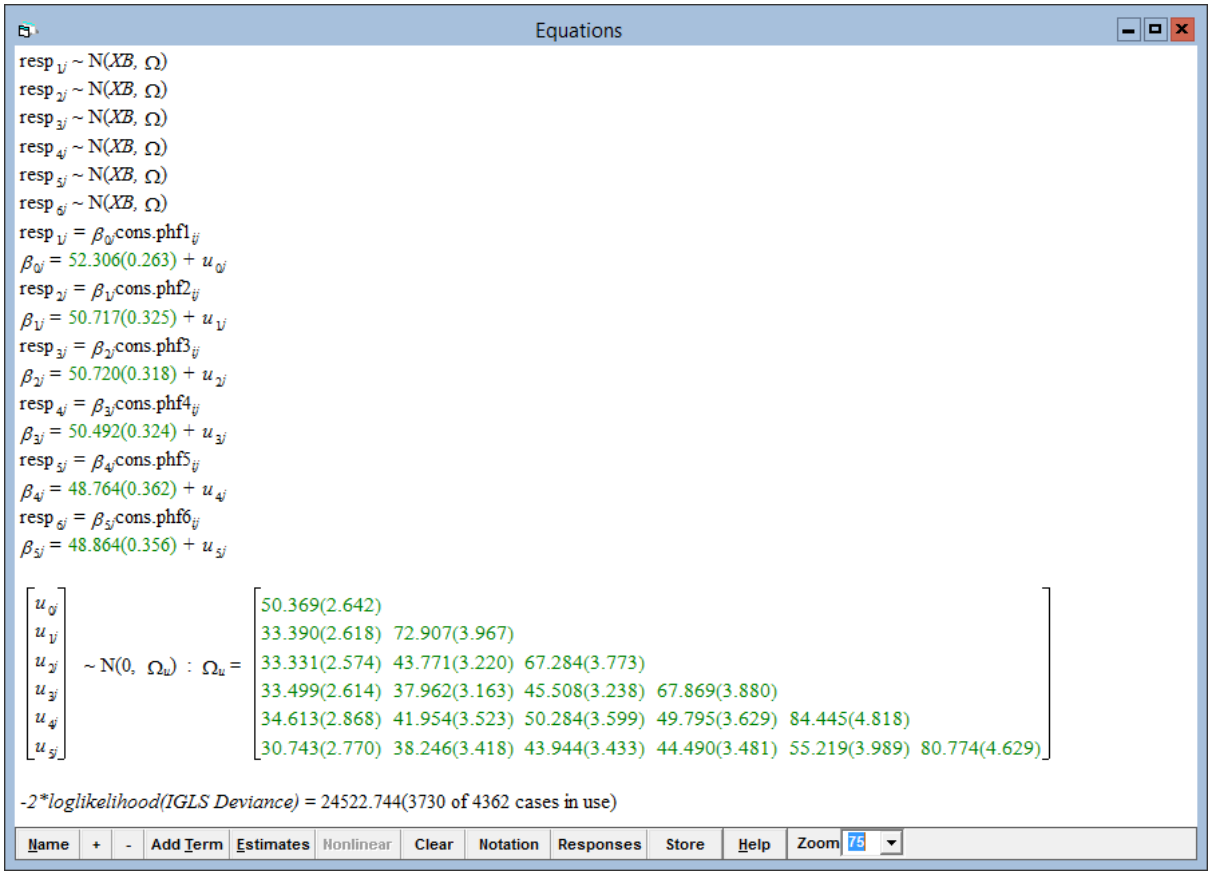
- From the **Data Manipulation menu, select Generate vector**
- In the **Output column** drop-down box select **c11**
- Press **Ctrl+N** to bring up the **Rename** window, enter **cons** and click **Rename**
- In the **Number of copies** enter **727** (the number of individuals in the data)
- In the **Value** box enter **1**
- Click **Generate**

In our multivariate model we enter the variables phf1 to phf6 as six separate response variables. Thus, the model has six separate equations. We include an intercept in each equation (**cons**) and a residual. We allow the six residuals to covary.

- From the **Model menu, select Equations**
- Click the **Responses** button at the bottom of the **Equations** window
- Select the response variables **phf1**-**phf6** and then click **Done**
- Click on **resp1** and change **N levels** to **2 – ij**, **level 2(j)** to **idwide**, **level 1(i)** to **resp_indicator**, and click **done**
- Click **Add Term** and select **cons** from the **variable** drop-down box then click **add separate coefficients** to estimate a separate intercept in each equation
- Click on each of the **cons** terms and check the **j(idwide_long)** boxes for each to enter a separate residual for each equation
- Click on **+** twice
- Click on **Estimates**
- Check that the window matches that shown below, then click **Start**

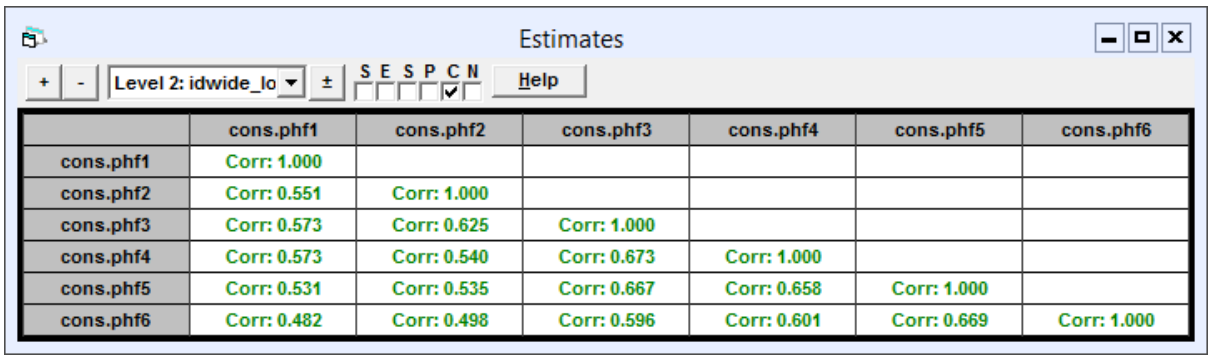


- Click **Estimates** a second time to reveal the parameter estimates and standard errors



The Equations window displays the estimated covariance matrix between the response occasions. To instead obtain the estimated correlations we use the **Estimates tables** window:

- From the **Model menu, select Estimate Tables**
- In the drop-down box at the top select **Level 2: idwide_long**
- Uncheck the four boxes to the right of the drop-down box and check the box labelled **C** for correlations



	cons.phf1	cons.phf2	cons.phf3	cons.phf4	cons.phf5	cons.phf6
cons.phf1	Corr: 1.000					
cons.phf2	Corr: 0.551	Corr: 1.000				
cons.phf3	Corr: 0.573	Corr: 0.625	Corr: 1.000			
cons.phf4	Corr: 0.573	Corr: 0.540	Corr: 0.673	Corr: 1.000		
cons.phf5	Corr: 0.531	Corr: 0.535	Corr: 0.667	Corr: 0.658	Corr: 1.000	
cons.phf6	Corr: 0.482	Corr: 0.498	Corr: 0.596	Corr: 0.601	Corr: 0.669	Corr: 1.000

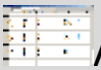
The estimated coefficients of the occasion dummies are close, but not exactly equal to, the sample means at each occasion. Similarly, there are some differences between the estimated correlations and the corresponding descriptive statistics. The differences we observe are due to missing data. In particular while the

descriptive statistics correlation matrix was based on the subset of individuals with no missing data, the model-implied correlation matrix is based on all available data.

P15.5 Adding Explanatory Variables: Fitting Group-specific Growth Curves

To open the worksheet:

From within the LEMMA Learning Environment

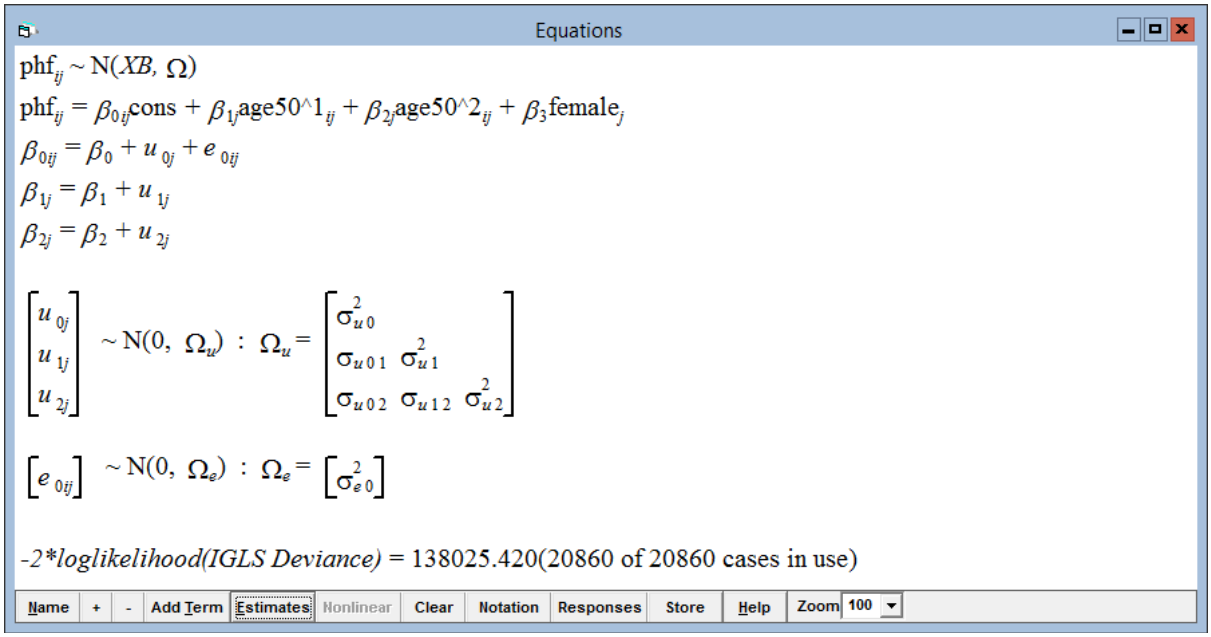
- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to  ***MlwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "  15.5.wsz"

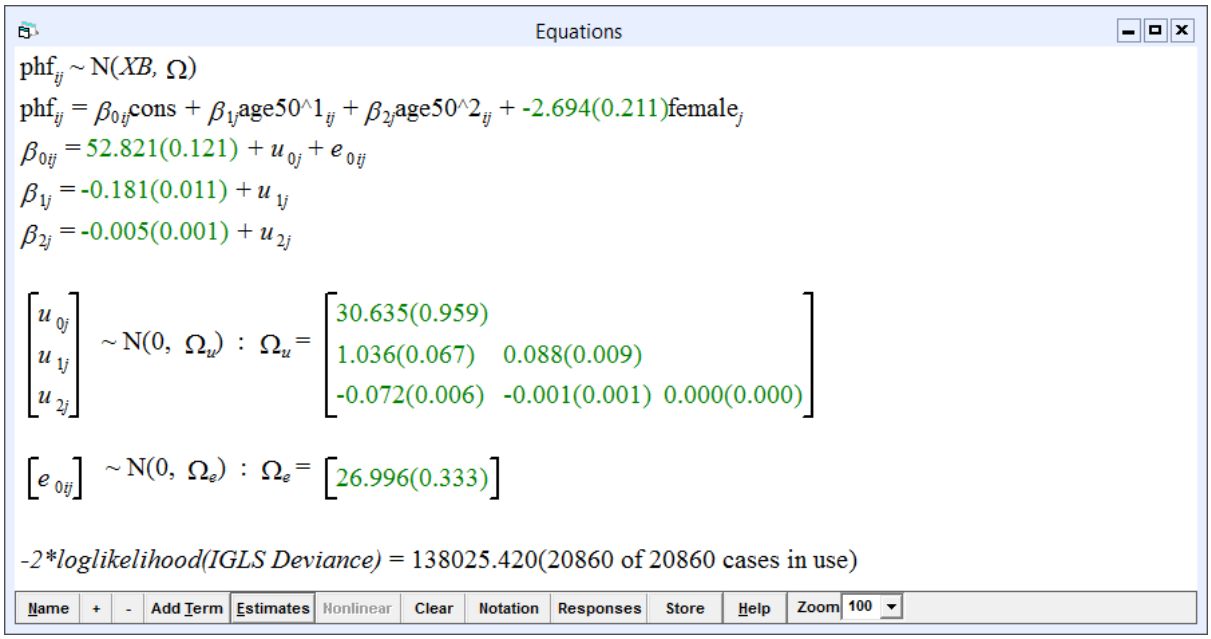
In the growth curve models considered thus far, the time variable is the only explanatory variable. The models can be extended to include several predictors, and these may be continuous and categorical. A common question in growth curve analysis is the extent to which the initial level of a response and its growth rate differ across sub-groups in the population. Such questions may be investigated by defining a categorical explanatory variable where the groups of interest form the categories. In this exercise, we illustrate the application of models that allow for differences between groups in an analysis of gender differences in physical health functioning.

We take as our starting point the quadratic model fitted at the end of P15.4.1. We begin by allowing for a gender difference in mean functioning at any age by including the dummy variable **female** (coded 1 for women and 0 for men) as an explanatory variable.

- From the **Model menu, select Equations**
- Click on the red **y**, select **phf** in the **y**: drop-down box, **2-ij** in the **N levels** drop-down box, **id** in **level 2(j)**, **occ** in **level 1(i)**, then click **done**
- Click on the red $\boldsymbol{\beta_0}$ term, select **cons** from the drop-down box, check the **j(id)** and **i(id)** boxes, and click **Done**
- Click **Add Term**, select **age50** from the **variable** drop-down box, check the **polynomial** check box, under **poly degree** select **2**, then click **Done**
- Click on **age50^1** and check the **j(id)** check box to add the first random slope
- Click on **age50^2** and check the **j(id)** check box to add the second random slope
- Click **Add Term**, select **female** from the **variable** drop-down box and then click **Done**
- Click the **+** button twice to see the full model specification

- Check that your model looks the same as that in the first screenshot below, click **Start**, and click the **Estimates** button twice to reveal the parameter estimates and standard errors

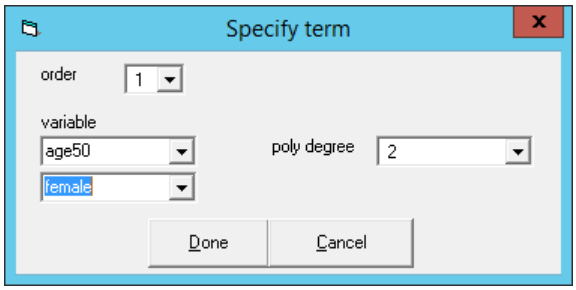


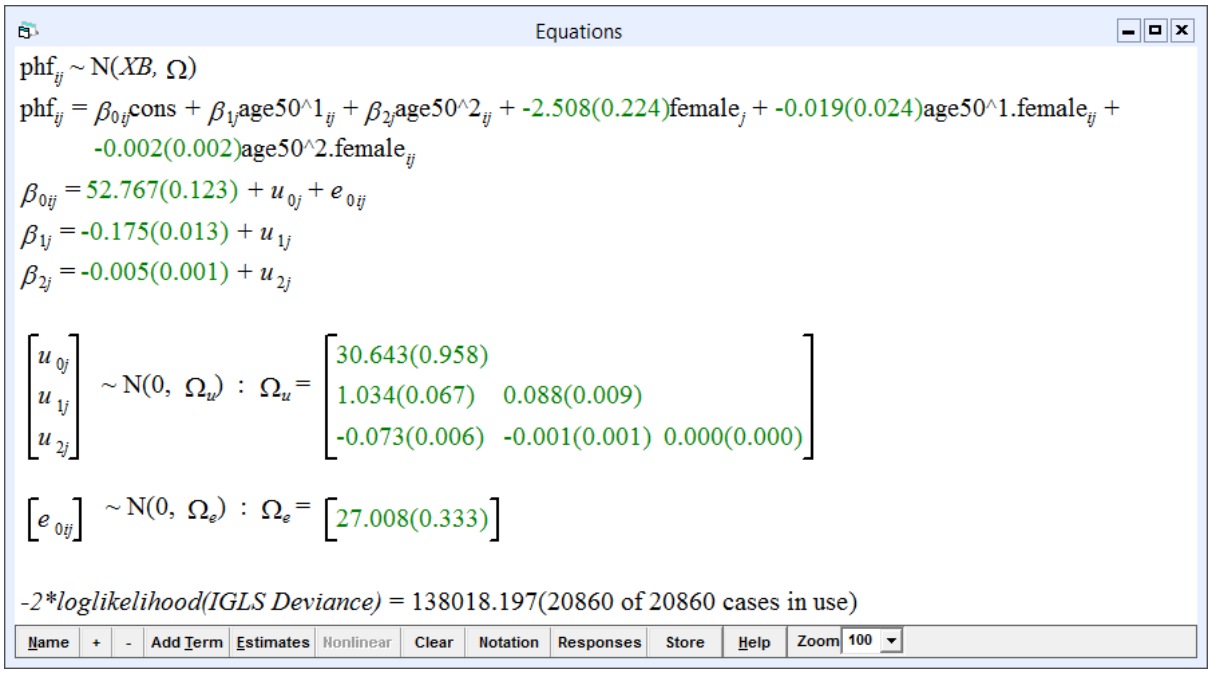


The gender effect is strongly significant, with women scoring 2.7 points lower than men on average. We next allow the effect of age on functioning (the rate of decline) to differ for men and women by including interactions between **female** and the age variables (**age50** and **age50_2**).

- Click **Add Term**, select **1** from the **order** drop-down box to request a first-order interaction (i.e., an interaction term between two variables), select **age50** from the first **variable** drop-down box, **female** from the second **variable** drop-down box, and click **Done**

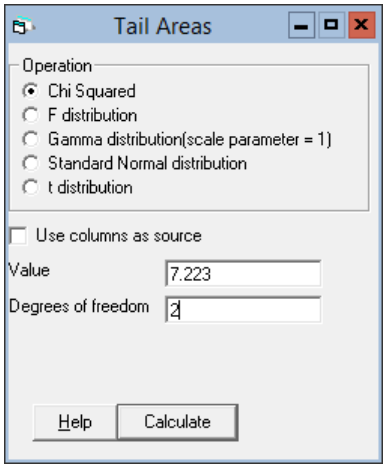
- Click **Start**





We compare the two models above using a likelihood ratio test. The difference between the model deviances is 7.223 (138025.420 - 138018.197). We can calculate the associated *p*-value using the **Tail areas** window.

- From the **Basic Statistics** menu, select **Tail Areas**
- Enter the value of **7.223** into the **Value** box
- Enter the value of **2** into the **Degrees of freedom** box
- Check that the window matches that shown below, then click **Calculate**




```
->CPRObability 7.223 2

 0.027011
```

The null hypothesis for the test is that the coefficients of the interaction variables **age50^1.female_1** and **age50^2.female_1** are both equal to zero. The null is rejected at the 5% level, so we conclude that the rate of decline in physical functioning with age differs for men and women. The average prediction curves for men and women are as follows.

Men (**female**=0): $\text{phf}_{ij} = 52.77 - 0.175\ \text{age50}_{ij} - 0.005\ \text{age50}_{ij}^2$

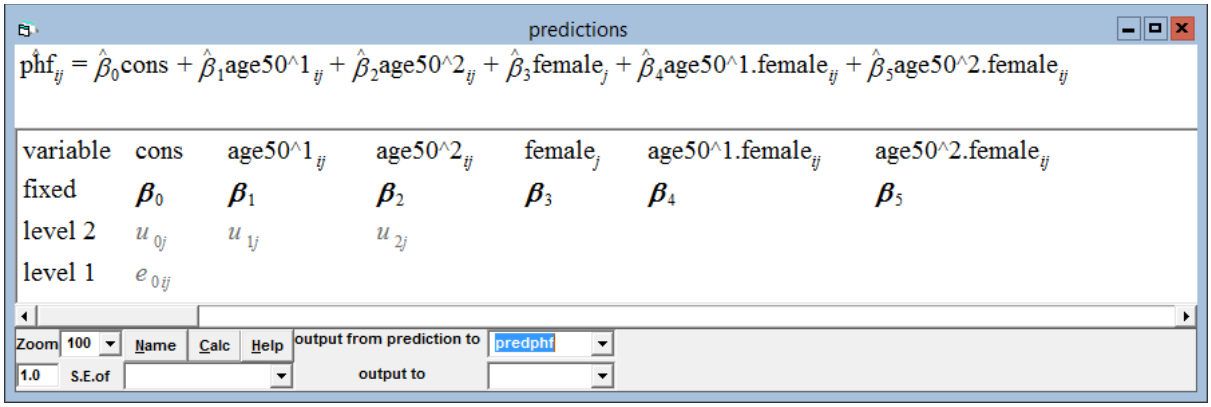
Women (**female**=1): $\text{phf}_{ij} = 50.26 - 0.194\ \text{age50}_{ij} - 0.007\ \text{age50}_{ij}^2$

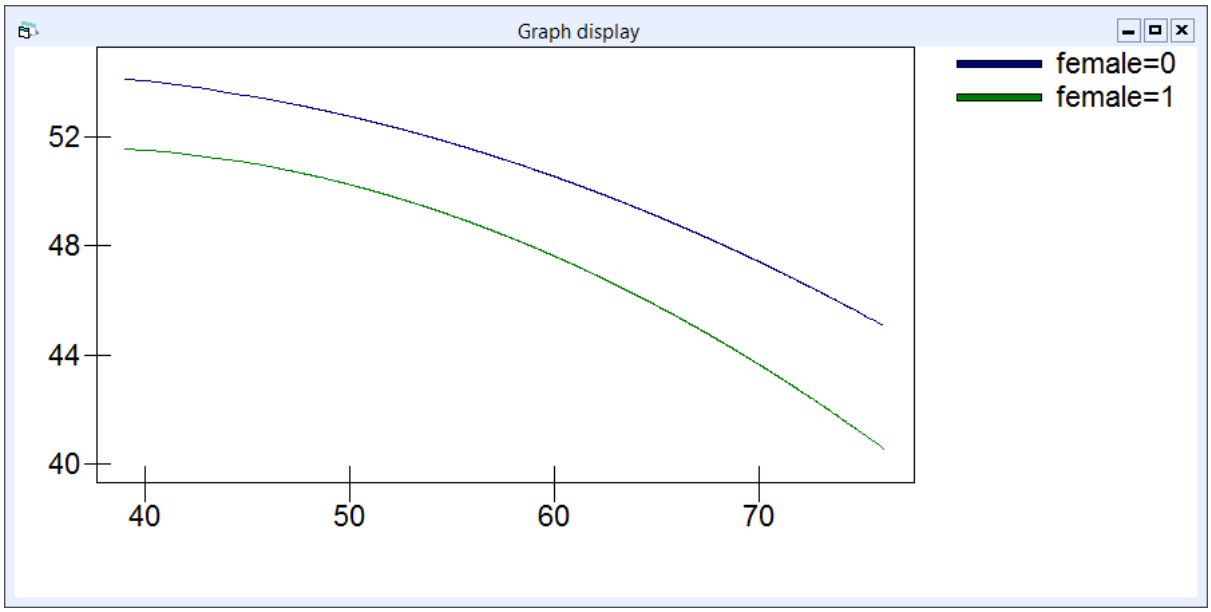
As the coefficients for both **age50^1** and **age50^2** are more negative for women than for men, the rate of decline is on average slower for women. We can visualise the gender difference by plotting the above functions using MLwiN's **Predictions** and **Customised graph** window. We calculate predicted values of **phf** and store them in a new variable called **predphf** (using the fixed part of the model only or, equivalently, setting the intercept and slope random effects to zero). We then sort by **age**[8] (the X-axis variable in the plot) and superimpose line graphs of **predphf** by **age** for women and men on the same plot.

- From the **Model** menu, select **Predictions**
- Click on **fixed** and then **Include all fixed coefficients** so that the **Predictions** window matches that shown below
- Select **c15** from the **output from prediction to** drop-down box
- Press **Ctrl + N** to bring up the **rename** window, enter **predphf** and click **Rename**
- Click **Calc**
- From the **Graphs** menu, select **Customised Graph(s)**
- In the **plot what?** tab select **predphf** as the **y** variable and **age** as the **x** variable, select **female** as the **group** variable so that separate trajectories are plotted for each gender, change **plot type** to **line**

[8] We could have used **age50** in place of **age**.

- On the **Other** tab, check the **group code** check box to add a legend to the graph
- In the **plot style** tab change **colour** to **16 rotate** so that the male and female trajectories are plotted in different colours, and click **Apply**





From the plot, we see that the rate of decline in **phf** with **age** is very similar for men and women, even though the difference is statistically significant.

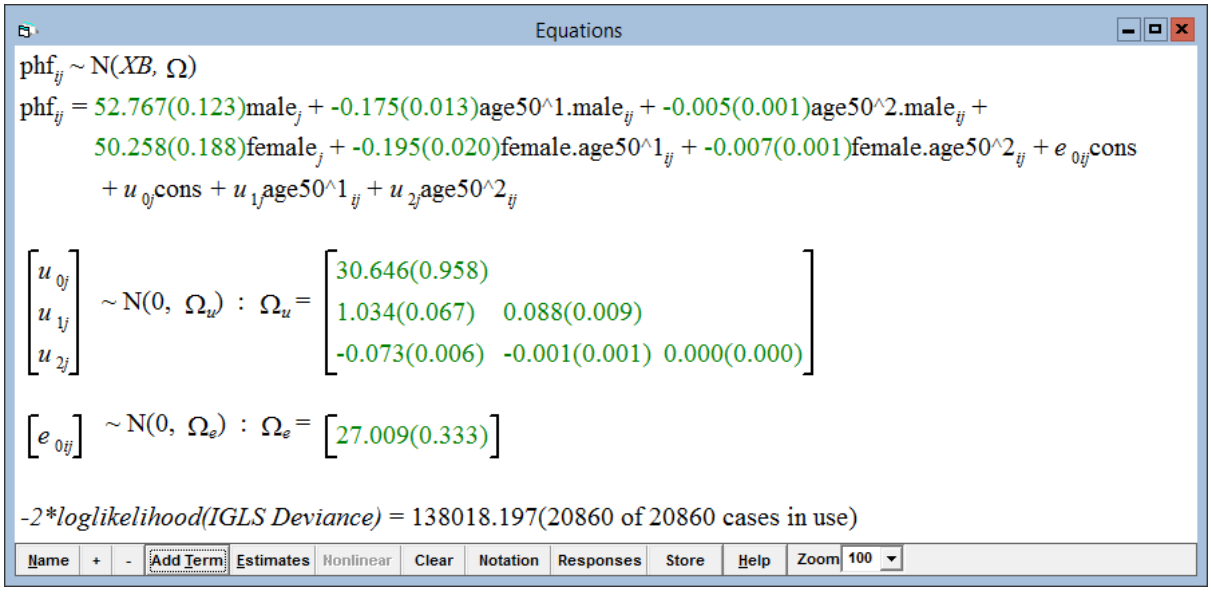
The model can be reparameterised to include dummy variables for both men and women and their interactions with the age variables. We calculate a dummy for men (called **male**) and interact this with **age50^1** and **age50^2**. These new variables are added to the model as explanatory variables, with the redundant constant removed.

- From the **Data Manipulation** menu, select **Command interface**, and enter the following two lines of code to generate a male dummy variable and to name it appropriately:

  CALC c15 = 1 - 'female'

  NAME c15 'male'

- In the **Equations** window click on the **cons** term, uncheck the **fixed Parameter** check box and click **Done** to remove it from the model
- Click on the **age50^1** term (i.e., the linear centred age term), uncheck the **fixed Parameter** check box and click **Done**
- Click on the **age50^2** term (i.e., the quadratic centred age term), uncheck the **fixed Parameter** check box and click **Done**
- Click on **female**, **age50^1.female** and **age50^2.female** (i.e., the old age by female interaction terms) and delete these terms
- Click **Add Term**, select the new **male** dummy variable from the **variable** drop-down box and click **Done** to add the male main effect
- Click **Add Term**, select **1** from the **order** drop-down box, select **age50** from the first **variable** drop-down box, **male** from the second **variable** drop-down box, and click **Done**. This will add the male by age interaction terms.
- Click **Add Term**, select **female** from the **variable** drop-down box and click **Done** to add the female main effect
- Click **Add Term**, select **1** from the **order** drop-down box, select **age50** from the first **variable** drop-down box, **female** from the second **variable** drop-down box, and click **Done.** This will add the female by age interaction terms.
- Click the **Estimates** button twice and check that your model is specified as in the first screenshot below, then click **Start** and click the **Estimates** button to see the model parameter estimates and standard errors



Notice that the log-likelihood value is exactly the same as for the previous model. This is expected because the two models are equivalent. The advantage of this parameterisation is that we obtain the coefficients for the male and female curves directly, without having to add together coefficients.

The current model allows the mean of **phf** at age 50 and the change in **phf** with age to depend on gender, but the between-individual variance in the intercept and slope is assumed the same for men and women. We can extend the model to allow the intercept and slope variances and their covariance to depend on gender by fitting

gender-specific random effects. This leads to the model given by equation (15.10) in C15.5:

$$\begin{aligned}\text{phf}_{ij} &= \beta_{0j}\text{male}_j + \beta_{1j}\text{age50}_{ij}\text{male}_j + \beta_{2j}\text{age50}_{ij}^2\text{male}_j \\ &+\beta_{3j}\text{female}_j + \beta_{4j}\text{age50}_{ij}\text{female}_j + \beta_{5j}\text{age50}_{ij}^2\text{female}_j + e_{ij}\end{aligned}$$

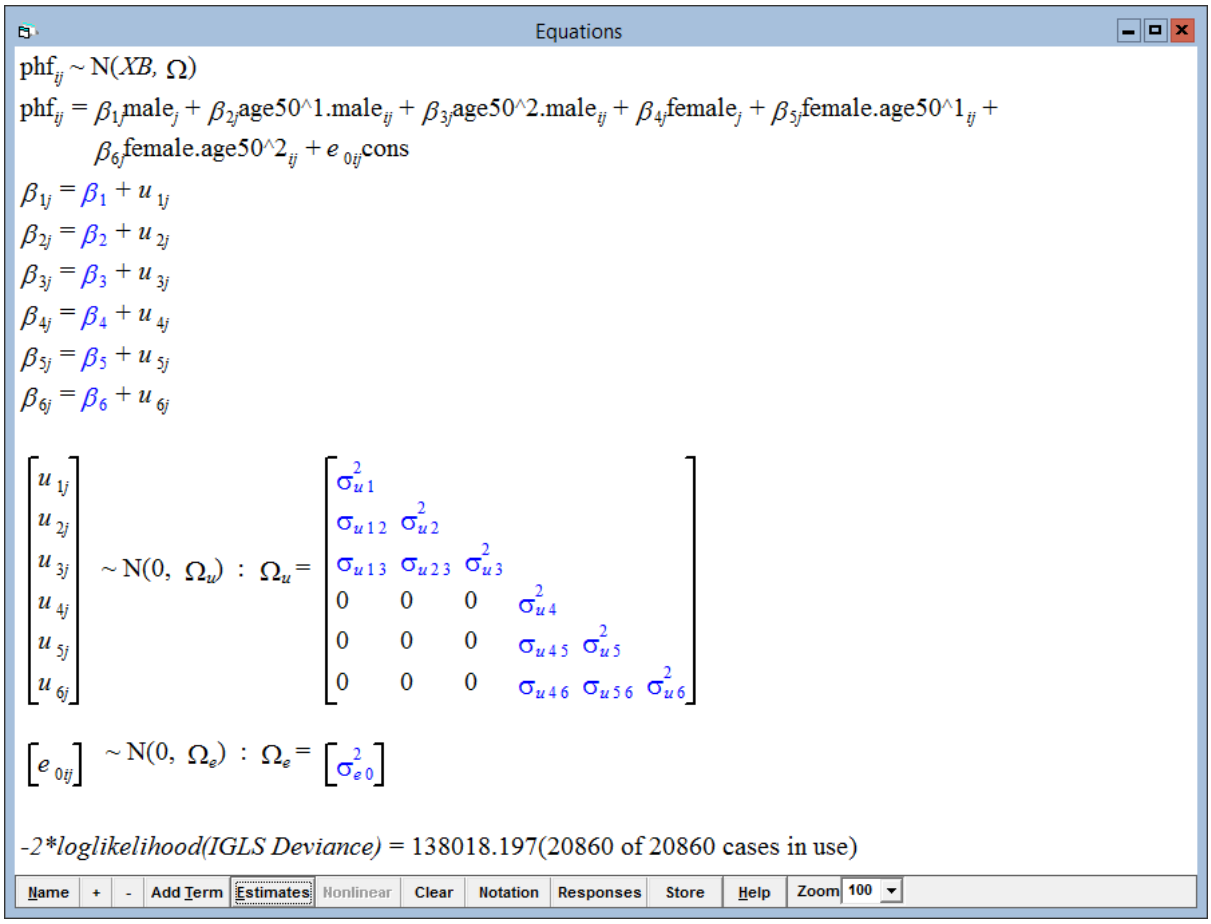
where $e_{ij} \sim N(0, \sigma_e^2)$, $\beta_{kj} = \beta_k + u_{kj}$ for $k = 0, 1, \ldots, 5$ and

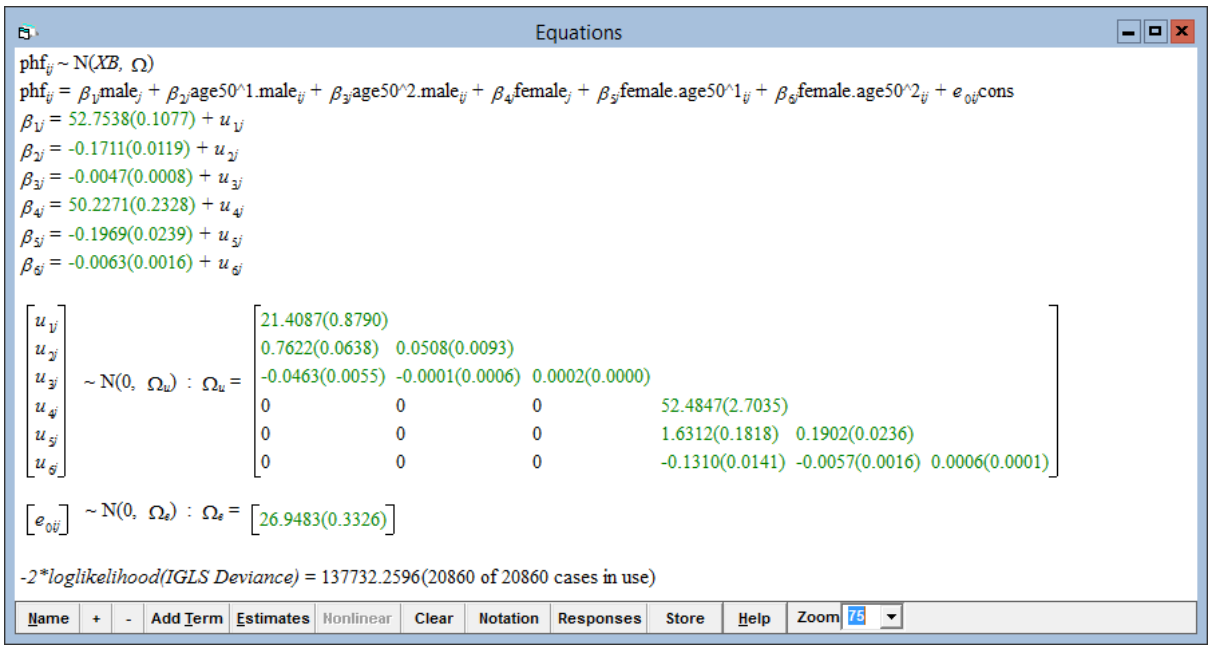
$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & & & & & \\ \sigma_{u01} & \sigma_{u1}^2 & & & & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 & & & \\ 0 & 0 & 0 & \sigma_{u3}^2 & & \\ 0 & 0 & 0 & \sigma_{u34} & \sigma_{u4}^2 & \\ 0 & 0 & 0 & \sigma_{u35} & \sigma_{u45} & \sigma_{u5}^2 \end{pmatrix}\right]$$

To fit this model we specify random coefficients for the **male**, **female**, and interaction terms[9]. The intercept random effects are the random coefficients for **male** and **female**, so the constants are redundant.

- In the **Equations** window click on the **cons** term then uncheck the **j(id)** check box and click **Done** to remove the common random intercept effect
- Click on the **age50^1** term, click **Delete Term** and then click **yes** in the box that pops up to remove the linear and quadratic terms for **age50**
- Click on each of the six remaining **male**, **female** and **age50** terms and check the **j(id)** box for each and click **Done**. This will add a random coefficient to every fixed-part covariate in the model.
- To remove the bottom left 3x3 sub matrix of covariance parameters, click on each of the these elements in turn and click **Yes** in the **remove term…** dialogue box that pops up
- Click the **Estimates** button twice and check that your model is specified as in the first screenshot below. The parameters to be estimated are highlighted in blue.
- Because some of the estimated covariance will be very small, we need to increase the display precision of the parameter estimates. To do this click on the **Options** menu, select **Worksheet…** , click the **Numbers** tab, and increase the **# digits after decimal point** to **4**. Click **Apply** and then **Done**
- In the **Equations** window click **Start** and click the **Estimates** button another time to reveal the parameter estimates and standard errors

[9] Note that in the Stata practical there were estimation difficulties and Stata was unable to fit this model. MLwiN does not suffer these same difficulties.





We compare the two models above using a likelihood ratio test. The difference between the model deviances is 285.93742 (138018.197 – 137732.2596). We calculate the corresponding *p*-value using the **Tail areas** window:

- From the **Basic Statistics** menu, select **Tail Areas**

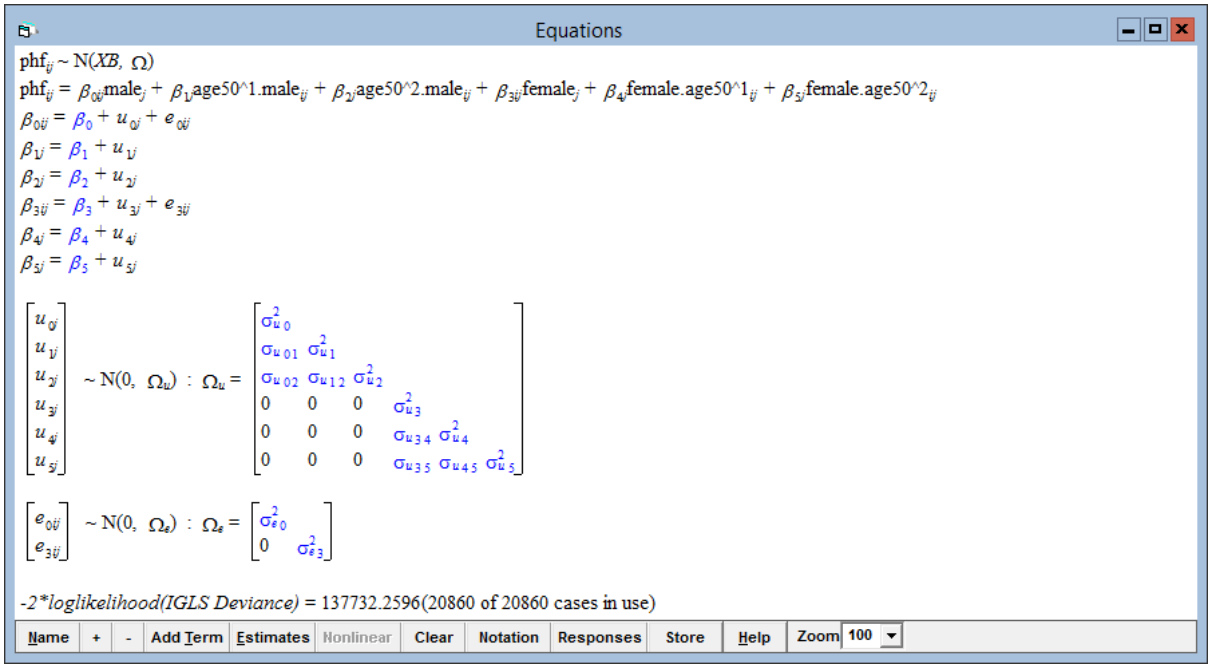
- Enter the value of **285.93742** into the **Value** box
- Enter the value of **6** into the **Degrees of freedom** box
- Click **Calculate**

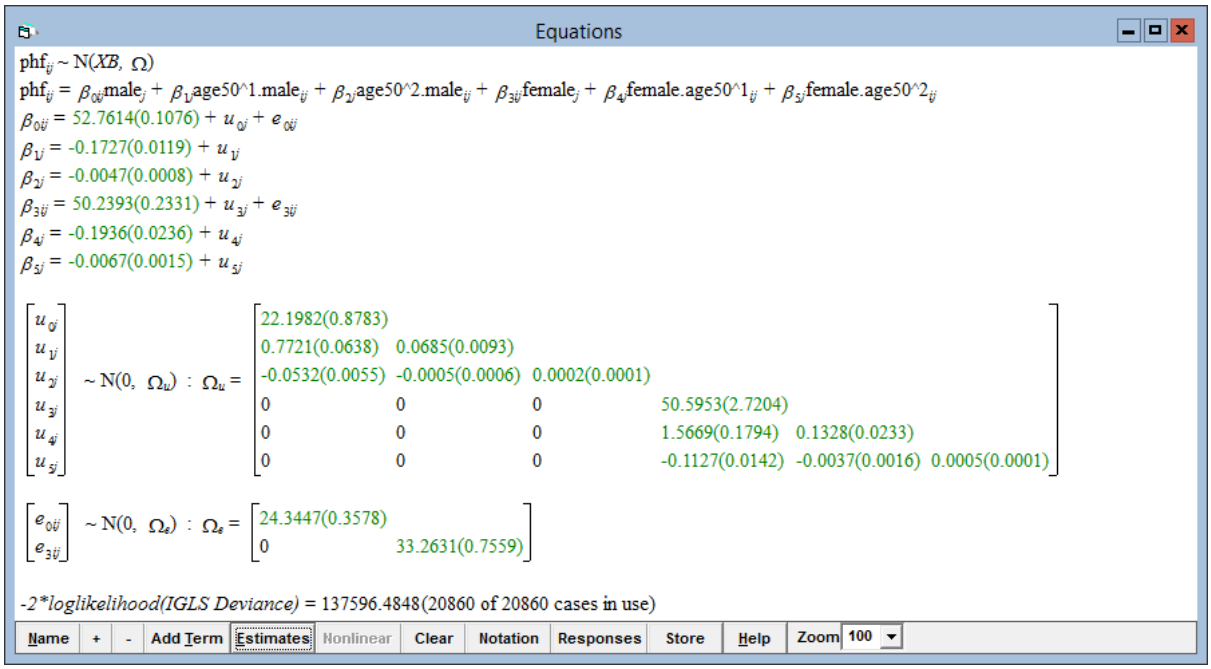
```
->CPRObability 285.93742 6

 0.00000
```

The null hypothesis of equal covariance structure is clearly rejected, so we conclude that the between-individual variance depends on gender. On examination of the estimates, we find that the between-individual variance at age 50 differs markedly for men and women: 21.41 for men and 52.48 for women. Physical health functioning at age 50 is substantially more variable for women than for men.

The final model we consider allows the within-individual variance to depend on gender. Do women exhibit more or less variation in health functioning with age than men? We adapt the model to fit different residual (within-individual) variances for groups defined by gender**.**

- In the **Equations** window click on the **cons** term, and click **Delete Term**. This removes the common residual variance which was imposed in the previous model.
- Click on the **female** term, check the **i(occ)** box and click **Done**. This adds a female specific residual.
- Click on the **male** term, check the **i(occ)** box and click **Done**. This adds a male specific residual.
- In the level 1 covariance matrix at the bottom of the model, click on the diagonal covariance term and select yes to the pop up box that says **remove term male/female_1 from level 1 covariance matrix?**
- Click **Estimates** and check that your model represents the first screenshot below
- Then click **Start** to run the model and then click **Estimates**





We carry out a likelihood ratio test of the null hypothesis that the within-individual variance is equal for men and women. The difference between the deviance of the current and previous model is 135.7748 (137732.2596 - 137596.4848). We can calculate the *p*-value in the usual way.

- From the **Basic Statistics** menu, select **Tail Areas**
- Enter the value of **135.77478** into the **Value** box
- Enter the value of **1** into the **Degrees of freedom** box
- Click **Calculate**

```
->CPRObability 135.77478 1

 0.00000
```

There is strong evidence that the within-individual variance differs for men and women at 24.34 for men and 33.26 for women. The within-individual variance measures the variability across measurement occasions in the deviations of individuals' observed physical health functioning scores from their quadratic growth curve. The higher residual variance for women suggests that their physical functioning trajectories are less well captured by a quadratic function than for men, even though the model allows for individual-specific linear and quadratic terms for women. Put another way, the fluctuations in health scores from occasion to occasion are greater for women than for men.

P15.6 Residual Autocorrelation

There is no practical for this section.

PART II: DYNAMIC (AUTOREGRESSIVE) MODELS

P15.7 Introduction to Dynamic Models

To open the worksheet:

From within the LEMMA Learning Environment
- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "15.7.wsz"

P15.7.1 Introduction to the smoking dataset

The application of random effects dynamic models will be illustrated using annual panel data on smoking behaviour. The response variable is the self-reported number of cigarettes smoked per day. The distribution of cigarettes smoked for the whole sample is highly skewed due to the high proportion of zeros contributed by non-smokers. For the purposes of illustration, we analyse the sub-sample of smokers for whom the logarithm of the number of cigarettes smoked is approximately normal.[10]

After excluding non-smokers, there are 5475 individuals in the analysis sample. Each individual is observed for up to 12 years between 1991 and 2002. The data are in long form with 27360 records, so the mean number of observations per individual is 5. The analysis file contains the following variables:

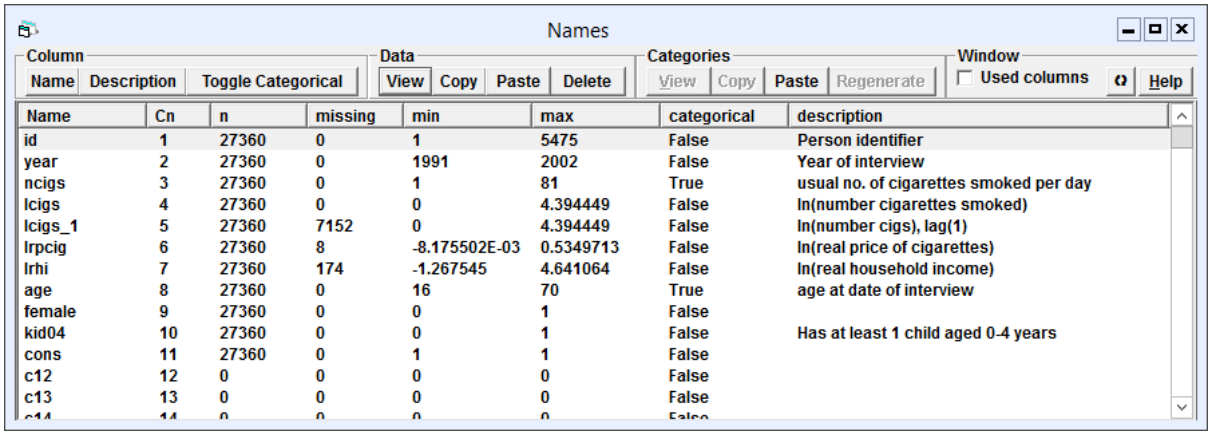
Variable	Description and codes
id	Individual identifier (coded 1, 2, . . ., 5475)
year	Year of interview (1991, . . ., 2002)
ncigs	Usual number of cigarettes smoked per day
lcigs	log(number of cigarettes smoker per day)*
lcigs_1	Lagged lcigs, i.e. reported number of cigarettes smoked when interviewed in the previous year
lrpcig	log(real price of cigarettes), based on a weighted average of the price across brands where weights are market shares

[10] A preferable approach would be to specify a model with two components: (i) a binary response model for whether an individual is a smoker, and (ii) a model for the number of cigarettes smoked among smokers. See Madden (2008) for a discussion of the use of sample selection models and two-part models in the analysis of cross-sectional data on tobacco and alcohol consumption. It is possible to extend both types of model to include random effects. For example, Steele and Durrant (2011) describe a multilevel sample selection model for survey nonresponse, separating the processes of contact and participation (conditional on contact).

lrhi	log(real household income)
age	Age at interview (years)
female	Gender (1=female, 0=male)
kid04	Respondent has at least one child age 0-4 years (1=yes, 0=no)

*Natural logarithms are used for all log-transformed variables.

We can confirm the contents of the MLwiN worksheet via inspecting the **Names** window.



Name	Cn	n	missing	min	max	categorical	description
id	1	27360	0	1	5475	False	Person identifier
year	2	27360	0	1991	2002	False	Year of interview
ncigs	3	27360	0	1	81	True	usual no. of cigarettes smoked per day
lcigs	4	27360	0	0	4.394449	False	ln(number cigarettes smoked)
lcigs_1	5	27360	7152	0	4.394449	False	ln(number cigs), lag(1)
lrpcig	6	27360	8	-8.175502E-03	0.5349713	False	ln(real price of cigarettes)
lrhi	7	27360	174	-1.267545	4.641064	False	ln(real household income)
age	8	27360	0	16	70	True	age at date of interview
female	9	27360	0	0	1	False	
kid04	10	27360	0	0	1	False	Has at least 1 child aged 0-4 years
cons	11	27360	0	1	1	False	
c12	12	0	0	0	0	False	
c13	13	0	0	0	0	False	

We obtain summary statistics for all variables using the **Averages and Correlation** window.

- From the **Basic Statistics** menu, select **Averages and Correlation**
- Select all of the variables, and click **Calculate**

```
->AVERage 11  'id'   'year'   'ncigs'   'lcigs'   'lcigs_1'   'lrpcig'   'lrhi'   'age'
'female'   'kid04'   'cons'

                N     Missing    Mean         s.d.
id      27360         0     2350.1        1402.1
year      27360         0     1996.2        3.4470
ncigs     27360         0     15.473        8.3241
lcigs     27360         0     2.5531        0.69139
lcigs_1     27360      7152     2.6143        0.63714
lrpcig     27360         8     0.26715       0.17860
lrhi      27360       174     1.6429        0.65947
age      27360         0     38.299        14.236
female      27360         0     0.51930       0.49964
kid04     27360         0     0.14130       0.34834
cons      27360         0     1.0000        0.00000
```

For each individual, the dataset has already been reduced to years when they participated in the survey and reported being a smoker. Records where **ncigs** was missing for any other reason have also been excluded. Three variables have fewer than 27360 non-missing observations: **lcigs_1**, **lrpcig** and **lrhi**. We would expect the lagged response **lcigs_1** to have missing values because this variable is only defined if the individual was present and a smoker at the previous year. There are few missing values on the other two variables.

We start by looking at the number of smokers with a non-missing value for **ncigs** in any given year:

- From the **Basic Statistics** menu, select **Tabulate**
- In the **Columns** drop-down box select drop-down box select **year**
- Check the **Percentages of column totals** check box
- Click **Tabulate**

```
->TABUlate 2  'year'

              1991      1992      1993      1994      1995      1996      1997      1998      1999      2000      2001      2002    TOTALS

   N          2740      2485      2296      2316      2277      2385      2347      2262      2159      2081      2047      1965     27360
   %          10.0       9.1       8.4       8.5       8.3       8.7       8.6       8.3       7.9       7.6       7.5       7.2     100.0
```

Bearing mind that the total number of smokers across the full 12-year period is 5475, it is clear that relatively few contribute information at each wave. The number of individuals is highest in the first year and lowest in the last, a pattern which is consistent with attrition.

We next examine the distribution of the number of measurement occasions per individual. The first step is to count the number of values of **ncigs** per individual (**numocc**). We then tabulate this variable, selecting one record per individual to obtain the frequency distribution for individuals rather than person-waves (as in the preliminary analysis of the physical functioning data in P15.1.3).

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- Under the **Operation** drop-down box, select **count**
- Under the **On blocks defined** by drop-down box, select **id** so that the count operation is performed separately for each individual
- Under **Output columns** select **c12**. This column will contain the number of measurement occasions for each individual.
- Click **Add to action list**, and then click **Execute**
- From the **data Manipulation** menu, select **Unreplicate**
- Under the **Take first entry in blocks defined by** drop-down box select **id**
- Under **Input columns** select **c12**
- Under **Output columns** select **c13**
- Click **Add to action list**, and then click **Execute**
- From the **Basic Statistics** menu, select **Tabulate**
- Check the **Percentages of column totals** check box
- In the **Columns** drop-down box select drop-down box select **c13**
- Click **Tabulate**

```
->TABUlate 2  'c13'

                1         2         3         4         5         6         7         8         9        10        11        12    TOTALS

   N         1455       772       468       382       276       286       252       232       242       239       307       564      5475
   %         26.6      14.1       8.5       7.0       5.0       5.2       4.6       4.2       4.4       4.4       5.6      10.3     100.0
```

A large number of individuals have only one record but, provided that **lcigs_1** is non-missing, they still contribute information to the analysis. We therefore retain these individuals.

Before fitting any models, we check that the lagged response **lcigs_1** is missing whenever there is a gap between occasions due to missing data. In other words $\mathbf{lcigs\_1}_{ij}$ should be missing if $\mathbf{ncigs}_{i-1,j}$ is missing. Care must be taken when calculating lags for a dataset where waves with missing data have been excluded: the previous record may not refer to the previous year. To carry out this check, we create a variable called **gap** which indicates whether the difference between the values of **year** for the current record and the previous year is greater than 1 (**gap** = 1) or equal to 1 (**gap** = 0). **gap** will be missing for an individual's first record.

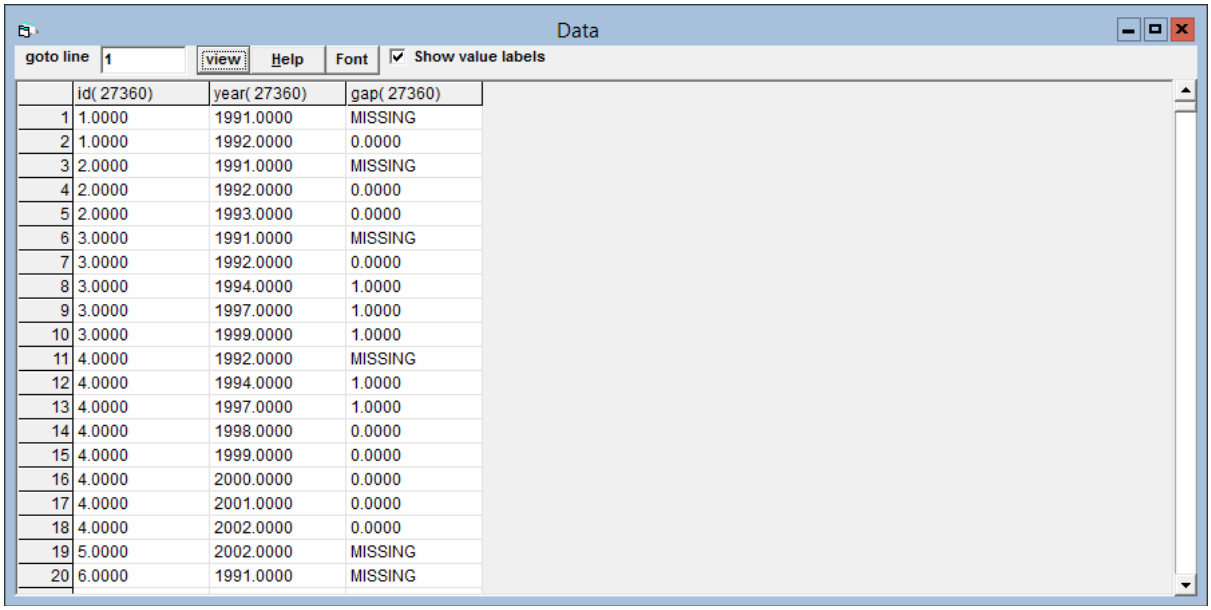
- From the **data Manipulation** menu, select **Multilevel data manipulations**
- Under the **Operation** drop-down box select **lags**
- Under the **On blocks defined** by drop-down box, select **id** so that the lag operation is performed separately for each individual
- Under **Input columns** select **year**
- Under **Output columns** select **c14**
- Click **Add to action list**, and then click **Execute**
- From the **Data Manipulation** menu, select **Command interface**, and enter the following lines of code to rename the lagged year variable stored in **c14** to **year_1** and to then generate a new variable named **gap**, stored in column **c15**, which is equal to **0** if there is no gap between occasions and **1** if there is

  NAME c14 'year_1'

  CALC c15=0*('year'=='year_1'+1)+1*('year'>'year_1'+1)

  NAME c15 'gap'
- In the **Names** window, select **id**, **year** and **gap**, then click **View** to view the new gap variable

The values of **id**, **year** and **gap** are given for the first 20 records below. Individual 1 has records only for 1991 and 1992 and therefore has no gaps. In contrast, individual 3 has several gaps due to missing data for years 1993, 1995, 1996 and 1998.



Data

goto line 1 | view | Help | Font | Show value labels

	id(27360)	year(27360)	gap(27360)
1	1.0000	1991.0000	MISSING
2	1.0000	1992.0000	0.0000
3	2.0000	1991.0000	MISSING
4	2.0000	1992.0000	0.0000
5	2.0000	1993.0000	0.0000
6	3.0000	1991.0000	MISSING
7	3.0000	1992.0000	0.0000
8	3.0000	1994.0000	1.0000
9	3.0000	1997.0000	1.0000
10	3.0000	1999.0000	1.0000
11	4.0000	1992.0000	MISSING
12	4.0000	1994.0000	1.0000
13	4.0000	1997.0000	1.0000
14	4.0000	1998.0000	0.0000
15	4.0000	1999.0000	0.0000
16	4.0000	2000.0000	0.0000
17	4.0000	2001.0000	0.0000
18	4.0000	2002.0000	0.0000
19	5.0000	2002.0000	MISSING
20	6.0000	1991.0000	MISSING

We check that **lcigs_1** is missing when the previous observation was more than 1 year ago, and find that this is indeed the case.

- From the **Basic Statistics** menu, select **Tabulate**
- Click the **Means** button
- In the **Variate column** drop-down box select **lcigs_1**
- In the **Columns** drop-down box select **gap**
- Click **Tabulate**

```
->TABUlate  'lcigs_1'    'gap'

7152 missing value(s)


Variable tabulated is lcigs_1

              0          1     TOTALS

      N   20208          0      20208
  MEANS       2.614      *          2.614
   SD'S       0.637      *          0.637
```

We see that when **gap** is equal to **1** there are no observed values for **lcigs_1** (the mean and SD cannot be calculated and are therefore reported as *). Hence **lcigs_1** always takes a missing value when the previous observation was more than 1 year ago.

P15.7.2 A simple random effects dynamic model for smoking

We begin by fitting an empty model with the logarithm of the number of cigarettes per day (**lcigs**) as the response and only a constant in the fixed-part of the model. We then fit a dynamic model by adding the cigarette lag (**lcigs_1**) as an explanatory variable. Finally we add the following additional explanatory variables: log of household income (**lrhi**), log of price of cigarettes (**lrpcig**), gender (**female**), **age**

and having young children (**kid04**). All three models also include an individual-level random effect.

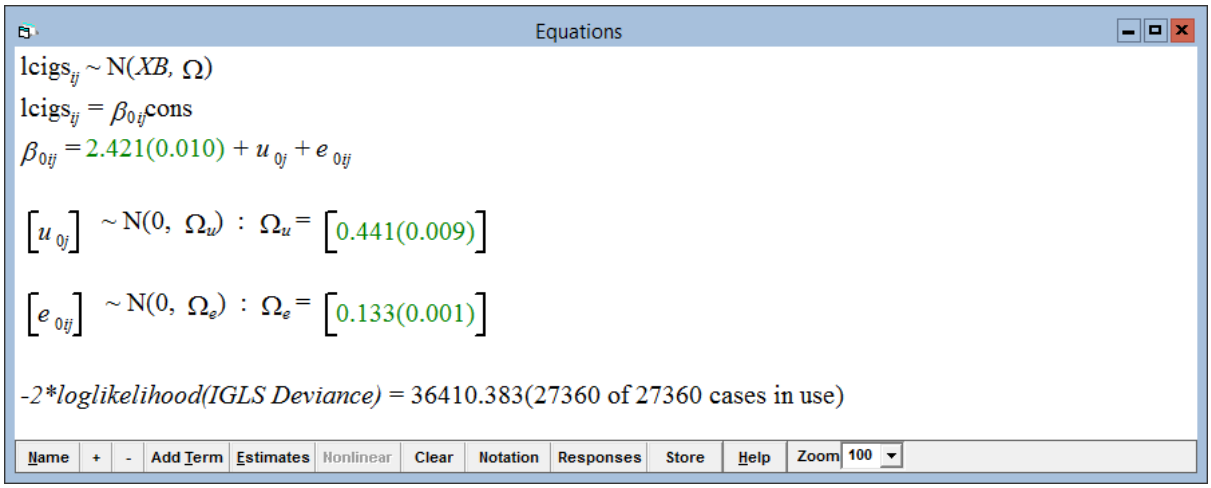
The full (third) model has the same form as equation (15.12) in C15.7.1

$$\text{lcigs}_{ij} = \beta_0 + \gamma \text{lcigs}_{i-1,j} + \beta_1 \text{lrhi}_{ij} + \beta_2 \text{lrpcig}_{ij} + \beta_3 \text{female}_j + \beta_4 \text{age}_{ij} + \beta_5 \text{kid04}_{ij} + u_j + e_{ij}$$

where $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.

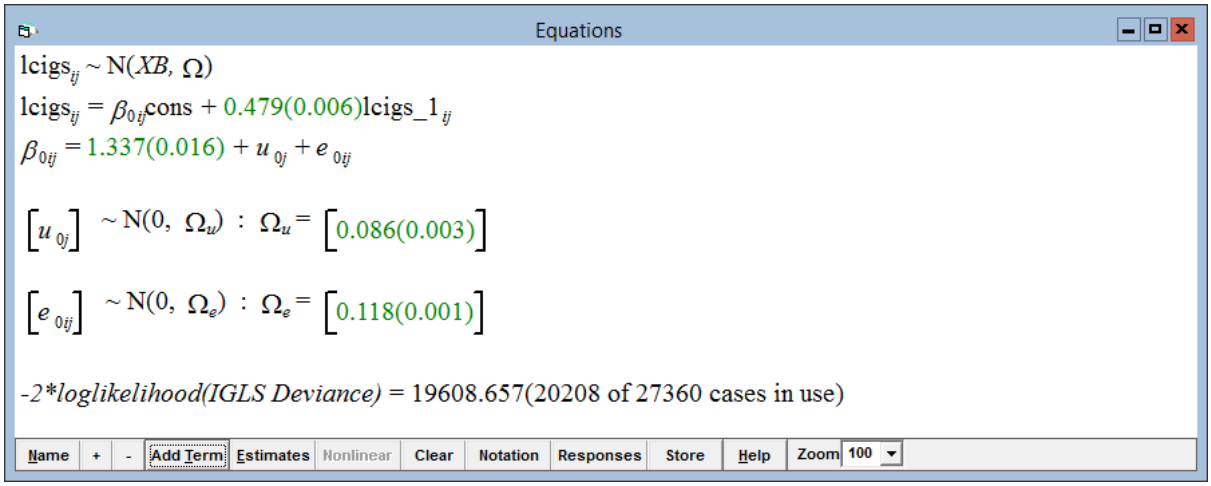
The above dynamic random effects model amounts to a random intercept multilevel model. We can set this up in the usual way.

- From the **Model** menu, select **Equations**
- Click the red **y**, select **lcigs** in the **y:** drop-down box, **2-ij** in the **N levels** drop-down box, **id** in **level 2(j)**, **year** in **level 1(i)**, then click **done**
- Click on the red $\boldsymbol{\beta_0}$ term, select **cons** from the drop-down box, check the **j(id)** and **i(id)** boxes, and click **Done**
- Click the **+** button twice to see the full model specification
- Click **Start**, and click the **Estimates** button twice to reveal the parameter estimates and standard errors
- Click **Store** at the bottom of the **Equations** window, name the model **model1**, and click **OK**

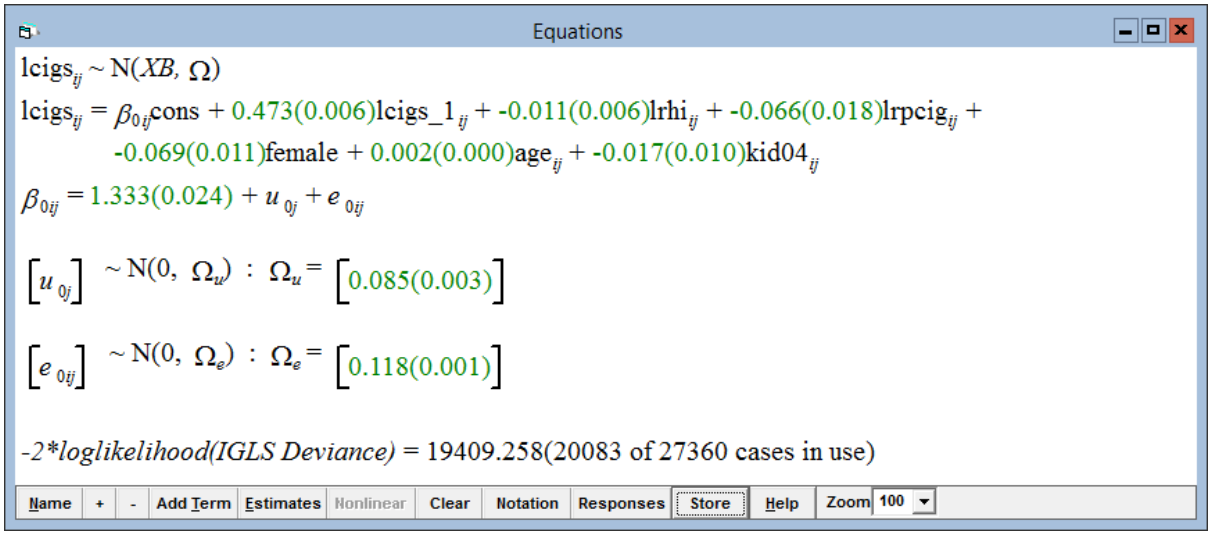


- Using the **Add Term** button, add the variable **lcigs_1** to the model
- Click **Start** to run the model
- Click **Store** and name the model **model2**

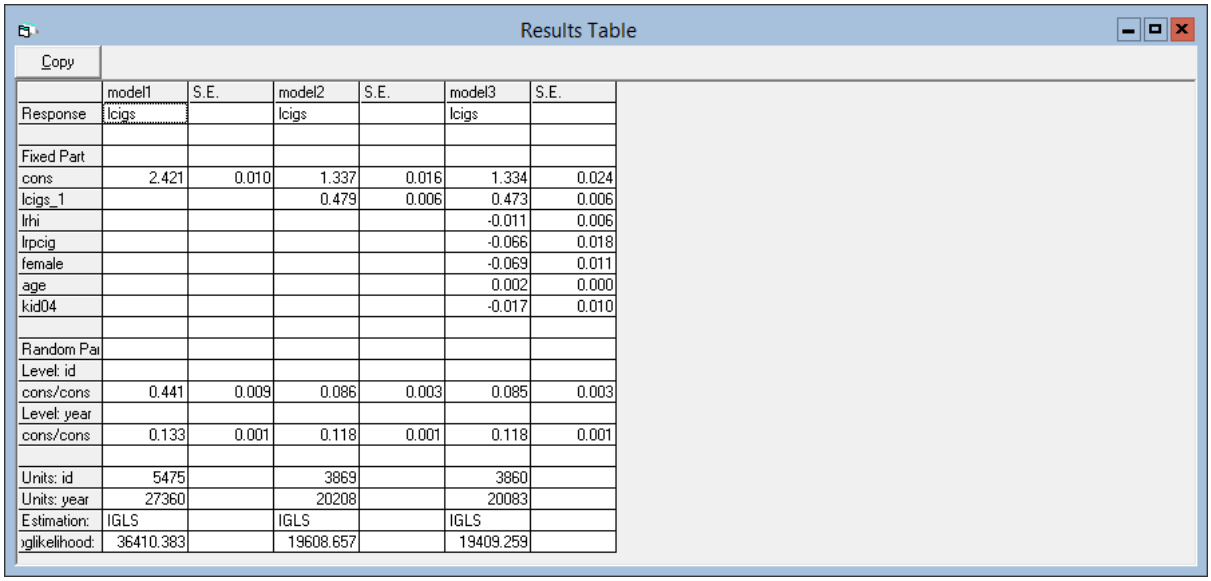
The model will take many iterations to converge, a common phenomenon in random effects dynamic models.



- Using the **Add Term** button, add the following explanatory variables to the model: **lrhi**, **lrpcig**, **female**, **age**, and **kid04**
- Click **More** to run the model using as starting values the parameter estimates from the previous model
- Click **Store** and name the model **model3**



- From the on **Model** menu, select **Manage stored models**
- Select all three models, **model1**, **model2** and **model3**
- Uncheck the **Correlation** check box on the right hand side of the window
- Click **Compare** to view the results of the three models side by side



Results Table

Copy

	model1	S.E.	model2	S.E.	model3	S.E.
Response	lcigs		lcigs		lcigs	
Fixed Part						
cons	2.421	0.010	1.337	0.016	1.334	0.024
lcigs_1			0.479	0.006	0.473	0.006
lrhi					-0.011	0.006
lrpcig					-0.066	0.018
female					-0.069	0.011
age					0.002	0.000
kid04					-0.017	0.010
Random Par						
Level: id						
cons/cons	0.441	0.009	0.086	0.003	0.085	0.003
Level: year						
cons/cons	0.133	0.001	0.118	0.001	0.118	0.001
Units: id	5475		3869		3860	
Units: year	27360		20208		20083	
Estimation:	IGLS		IGLS		IGLS	
loglikelihood:	36410.383		19608.657		19409.259	

The number of individuals (groups) in the third model is 3860 rather than the total number in the data file of 5475. The large reduction in sample size is due to a high proportion of those with only 1 record having a missing value for **lcigs_1**.

The coefficient of the lagged response γ is estimated as 0.473 (with a standard error of 0.006). As logarithms have been taken of both the response and its lag, we can transform γ to obtain a more interpretable quantity which measures the expected percentage change in tobacco consumption at occasion i for a given percentage change in consumption at occasion $i-1$. For example, a 10% increase in consumption at $i-1$ is associated with an increase in consumption at i by a factor of $1.10^{0.47} = 1.05$, i.e. a 5% increase.[11] A 25% increase in consumption at $i-1$ is associated with a 11% increase in consumption at i (from $1.25^{0.47} = 1.11$).

Turning to unobserved heterogeneity, the random effect variance is 0.085 implying a standard deviation of 0.292. For a 1 standard deviation increase in the random effect u_j, we would expect tobacco consumption to increase by a factor of $\exp(0.292) = 1.34$, i.e. a 34% increase.

Based on the results from this model, we would conclude that there is strong evidence of both state dependence and unobserved heterogeneity. The dependence between a person's tobacco consumption over time is driven by a combination of (i) a causal effect of previous consumption, and (ii) unmeasured time-invariant factors influencing consumption at any year (such as tendency towards addictive behaviour).

[11] Consider a simple regression of the form $\log(y) = a + b\log(x)$. Suppose we want the effect of increasing the value of x from x_1 to x_2, and denote by $\log(y)(x_k)$ the expected value of $\log(y)$ at $x = x_k$. The expected change in $\log(y)$ is $\log(y)(x_2) - \log(y)(x_1) = b[\log(x_2) - \log(x_1)]$. This can be expressed as $\log[y(x_2)/y(x_1)] = b\log[x_2/x_1]$, where $y(x_k)$ is the expected value of y at $x = x_k$. Taking exponentials of each side gives $y(x_2)/y(x_1) = [x_2/x_1]^b$. For a 10% increase in $x_2/x_1 = 1.10$, and the expected ratio of y at x_2 to y at x_1 is 1.10^b.

See http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm for further details of the interpretation of coefficients from regression models where a response variable and/or an explanatory variable are log-transformed.

Cigarette price and household income are both negatively associated with tobacco consumption, women smoke less than men, and consumption is higher among older smokers than younger smokers. After controlling for the effects of these variables, the effect of having young children is not significant at the 5% level.

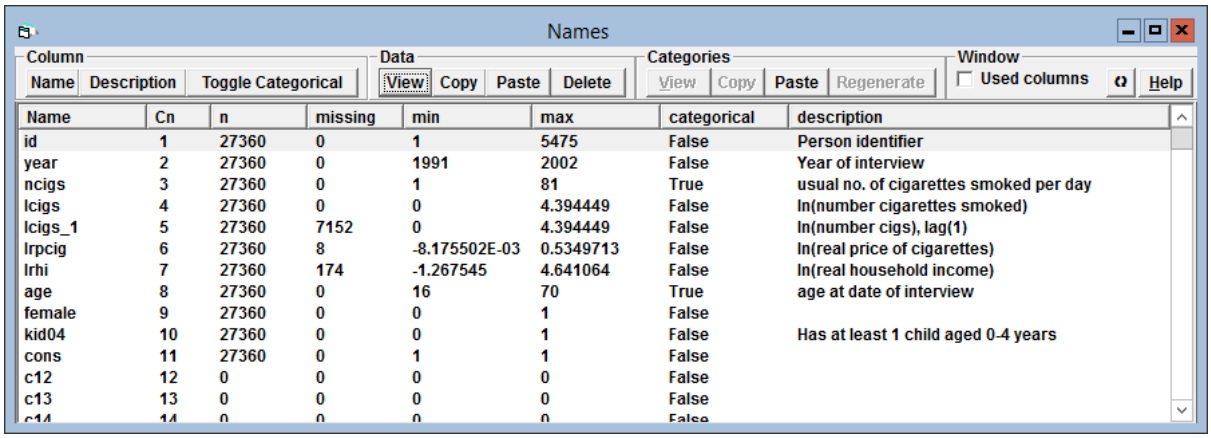
P15.8 The Initial Conditions Problem

To open the worksheet:

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to ***MLwiN* Datafiles**
- If you do not already have MLwiN to open the datafile with, click (**get *MLwiN***).
- Click "15.8.wsz"

The **Names** window should appear as follows:



Name	Cn	n	missing	min	max	categorical	description
id	1	27360	0	1	5475	False	Person identifier
year	2	27360	0	1991	2002	False	Year of interview
ncigs	3	27360	0	1	81	True	usual no. of cigarettes smoked per day
lcigs	4	27360	0	0	4.394449	False	ln(number cigarettes smoked)
lcigs_1	5	27360	7152	0	4.394449	False	ln(number cigs), lag(1)
lrpcig	6	27360	8	-8.175502E-03	0.5349713	False	ln(real price of cigarettes)
lrhi	7	27360	174	-1.267545	4.641064	False	ln(real household income)
age	8	27360	0	16	70	True	age at date of interview
female	9	27360	0	0	1	False	
kid04	10	27360	0	0	1	False	Has at least 1 child aged 0-4 years
cons	11	27360	0	1	1	False	
c12	12	0	0	0	0	False	
c13	13	0	0	0	0	False	
c14	14	0	0	0	0	False	

P15.8.1 Incorporating a model for smoking at occasion 1

The analysis in the previous exercise ignored the initial conditions problem. The lagged response **lcigs_1** was assumed to be uncorrelated with the individual random effect u_j. As explained in C15.8, this assumption is untenable because the dynamic model implies that $y_{i-1,t}$ also depends on u_j. The problem stems from treating tobacco consumption at an individual's first measurement occasion only as a predictor of consumption at the following wave. We would expect the omitted time-invariant variables represented by u_j to affect consumption at *all* occasions, including the first. In this exercise, we extend the model of P15.7 to include a model for the response **lcigs** at occasion 1, thereby allowing u_j to influence a person's consumption at each occasion.

We will consider models for **lcigs** at occasion 1 of the same form as equations (15.16) and (15.17) in C15.8.2. In our smoking example, the more general of the two models, given by equation (15.17), is specified as

$$\text{lcigs}_{1j} = \alpha_0 + \alpha_1 \text{lrhi}_{1j} + \alpha_2 \text{lrpcig}_{1j} + \alpha_3 \text{female}_j + \alpha_4 \text{age}_{ij} + \alpha_5 \text{kid04}_{1j} + \lambda u_j + e_{1j} \quad (A)$$

where $u_j \sim N(0, \sigma_u^2)$ as before and $e_{1j} \sim N(0, \sigma_{e1}^2)$.

The simpler model comes from setting $\lambda = 1$.

The model for lcigs at $i = 1$ is estimated jointly with the earlier model for lcigs at $i > 1$, i.e.

$$\text{lcigs}_{ij} = \beta_0 + \gamma\, \text{lcigs}_{i-1,j} + \beta_1 \text{lrhi}_{ij} + \beta_2 \text{lrpcig}_{ij} + \beta_3 \text{female}_j + \beta_4 \text{age}_{ij} + \beta_5 \text{kid04}_{ij} + u_j + e_{ij} \quad (B)$$

where $e_{ij} \sim N(0, \sigma_e^2)$.

To summarise, we estimate the following models which are labelled as in C15.8.

Model	Description
Model 1	Equation (B) only. This model ignores the initial conditions problem and was fitted in P15.7 and referred to there as model 3
Model 2	Joint model consisting of equation (A) with $\lambda = 1$ and equation (B)

P15.8.2 Fitting joint models in MLwiN

Model 2 can be fitted by creating two dummy variables for occasions $i = 1$ and $i > 1$ respectively, and interacting them with the explanatory variables in the equations for y_{1j} and y_{ij} for $i > 1$. To illustrate this approach, consider a simplified version of Model 2 where the lagged response (for $i > 1$) and household income (**lrhi**) are the only predictor variables. The models for $i = 1$ and $i > 1$ are

$$\text{lcigs}_{1j} = \alpha_0 + \alpha_1 \text{lrhi}_{1j} + u_j + e_{1j} \qquad (A1)$$

$$\text{lcigs}_{ij} = \beta_0 + \gamma\, \text{lcigs}_{i-1,j} + \beta_1 \text{lrhi}_{ij} + u_j + e_{ij} \qquad (B1)$$

We define two dummy variables as follows:

$$\mathbf{t1} = \begin{cases} 1 & \text{occasion 1 } (i = 1) \\ 0 & \text{otherwise } (i > 1) \end{cases}$$

$$\mathbf{tg1} = 1 - \mathbf{t1} = \begin{cases} 1 & \text{occasions } 2, 3, \ldots \text{T } (i > 1) \\ 0 & \text{otherwise } (i = 1) \end{cases}$$

t1 is interacted with **lrhi** because **lrhi** is the only predictor of **lcigs** at $i = 1$. Denote the interaction variable by **t1Xlrhi** = **t1** × **lrhi**.

tg1 is interacted with both the lagged response **lcigs_1** and **lrhi**, giving two interaction variables: **tg1Xlcig1** = **tg1** × **lcigs_1** and **tg1Xlrhi** = **tg1** × **lrhi**.

The joint model given by equations (A1) and (B1) above can be written in terms of these new variables as

$$\text{lcigs}_{ij} = \alpha_0\, t1_{ij} + \alpha_1 \text{t1Xlrhi}_{1j} + \beta_0 \text{tg1}_{ij} + \gamma\, \text{tg1Xlcig1}_{ij} + \beta_1 \text{tg1Xlrhi}_{ij} + u_j + e_{ij} \quad (C1)$$

where $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_{ei}^2)$ such that $\sigma_{ei}^2 = \sigma_e^2$ for $i > 1$.

Substitution of **t1**=1 and **tg1**=0 in equation (C1) gives equation (A1), while substitution of **t1**=0 and **tg1**=1 gives equation (B1). Thus the single-equation model of (C1) is equivalent to the joint model given by (A1) and (B1). This is a useful trick for fitting joint models in general.

One point to note is that the interaction with the lagged response **tg1Xlcig1** must be coded as zero for occasion $i = 1$. If **tg1Xlcig1** is coded as missing for occasion 1 (because the lag **lcigs_1** is missing), all observations for occasion 1 will be dropped from the analysis sample. Like other statistics packages, MLwiN uses listwise deletion to handle missing data on predictor variables whereby an entire record is deleted if there is a missing value for any predictor in the specified model.

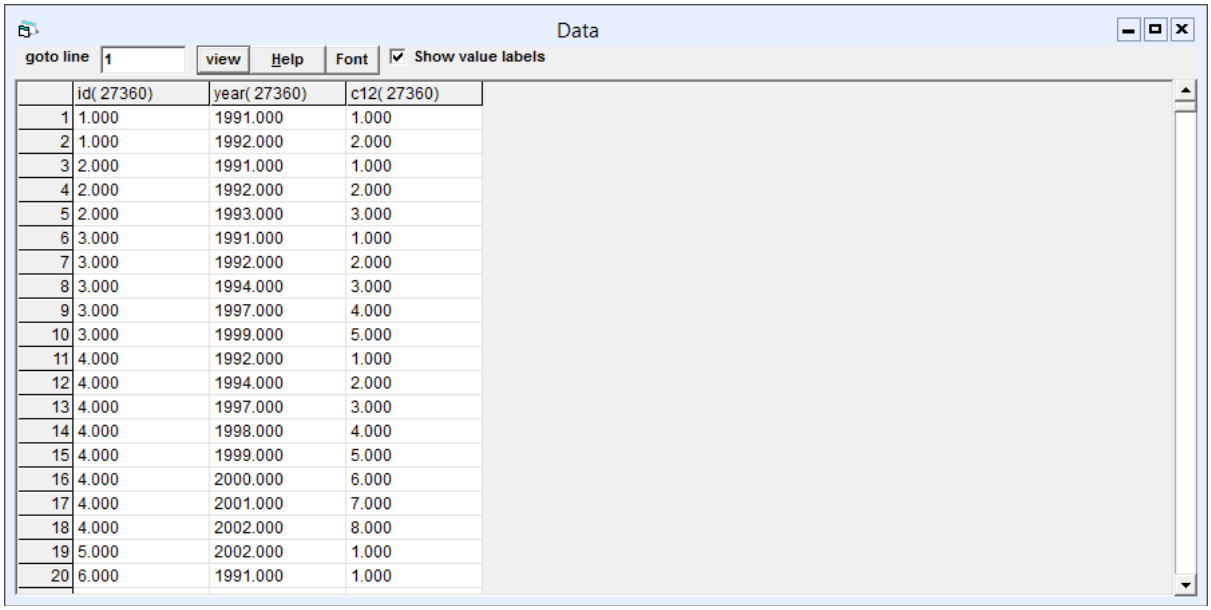
P15.8.3 Results

Model 2

We begin by fitting the full Model 2 given by equations (A) and (B) in P15.18.1 with $\lambda = 1$.

We first create the dummy variables for occasions:

- From the **Data Manipulation** menu, select **Multilevel data manipulations**
- In the **Operation** box select **Sequence**
- In the **On blocks defined by** box select **id**
- Select **c12** from **Output columns**
- Click **Add to action list**, and then **Execute**
- In the **Names** window, highlights **id**, **year** and **c12**, and then click **View** to confirm that the new variable **c12** indexes the observations within each individual



	id(27360)	year(27360)	c12(27360)
1	1.000	1991.000	1.000
2	1.000	1992.000	2.000
3	2.000	1991.000	1.000
4	2.000	1992.000	2.000
5	2.000	1993.000	3.000
6	3.000	1991.000	1.000
7	3.000	1992.000	2.000
8	3.000	1994.000	3.000
9	3.000	1997.000	4.000
10	3.000	1999.000	5.000
11	4.000	1992.000	1.000
12	4.000	1994.000	2.000
13	4.000	1997.000	3.000
14	4.000	1998.000	4.000
15	4.000	1999.000	5.000
16	4.000	2000.000	6.000
17	4.000	2001.000	7.000
18	4.000	2002.000	8.000
19	5.000	2002.000	1.000
20	6.000	1991.000	1.000

In order to fit interactions between occasion and the predictor variables, we must create two new dummy variables **t1** (for occasion 1) and **tg1** (for subsequent occasions). This is also necessary as we want to replace missing values for **tg1Xlcig** with zeros for only occasion 1.

- From the **Data manipulation** menu, select **Command interface**
- Run the following lines of code:

  CALC c13=(c12==1)
  CALC c14=(c12>1)
  NAME c13 ‘t1’
  NAME c14 ‘tg1’

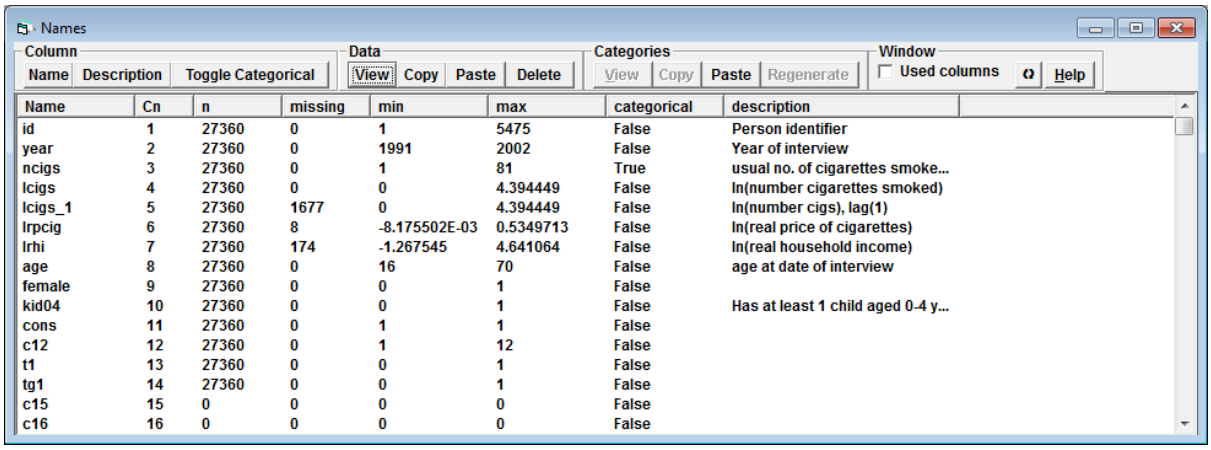
Gaps between occasions are another reason for missing values on the lagged response; we want to exclude these records from the analysis rather than recode to 0.

- From the **Data Manipulation** menu, select **recode** then select **by Range**
- Enter **missing** into the first **Values in range** box, **missing** into the second **Values in range** box and **-999** into the **to new value** box
- In **Input columns** select **lcigs_1** and in **Output columns** select **lcigs_1**
- Click **Add to action list**
- Click **Execute**
- From the **Data manipulation** menu, select **Command interface**
- Run the following lines of code:

  CALC ‘lcigs_1’ = ‘lcigs_1’ - ‘t1’
- From the **Data Manipulation** menu, select **recode** menu, select **by Range**
- Enter **-999** into the first **Values in range** box, **-999** into the second **Values in range** box and **missing** into the **to new value** box then
- In **Input columns** select **lcigs_1** and in **Output columns** select **lcigs_1**

- Click **Add to action list**
- Enter **-1000** into the first **Values in range** box, **-1000** into the second **Values in range** box and **0** into the **to new value** box then click **Add to action list**
- Click **Execute**

The **Names** window should now look like this:

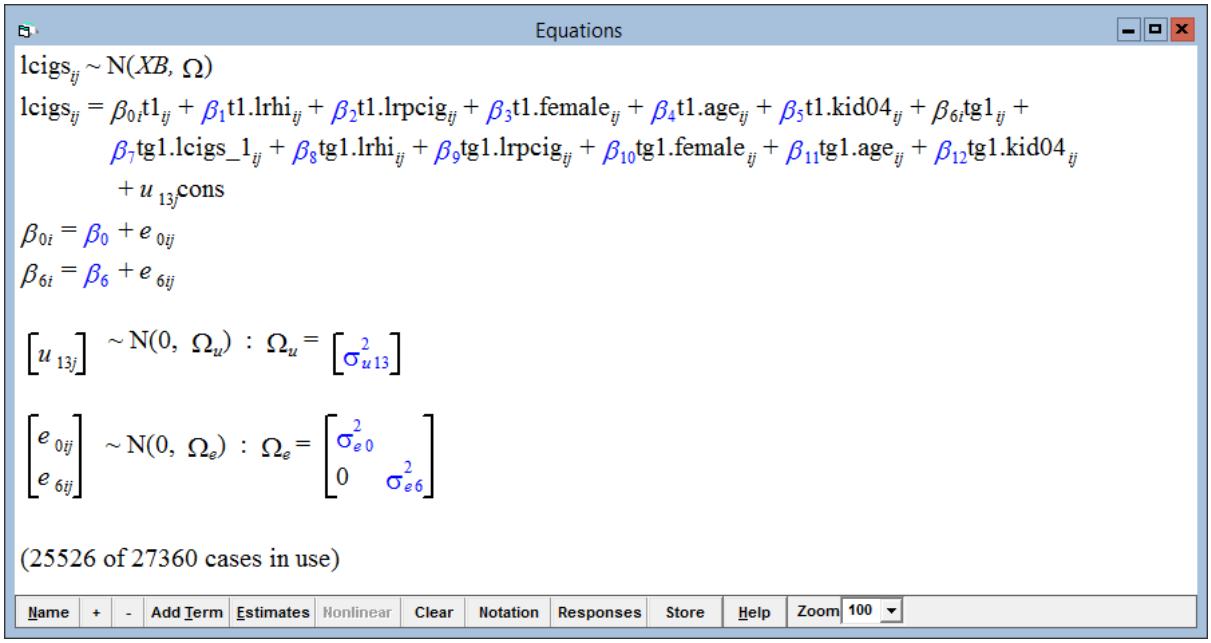


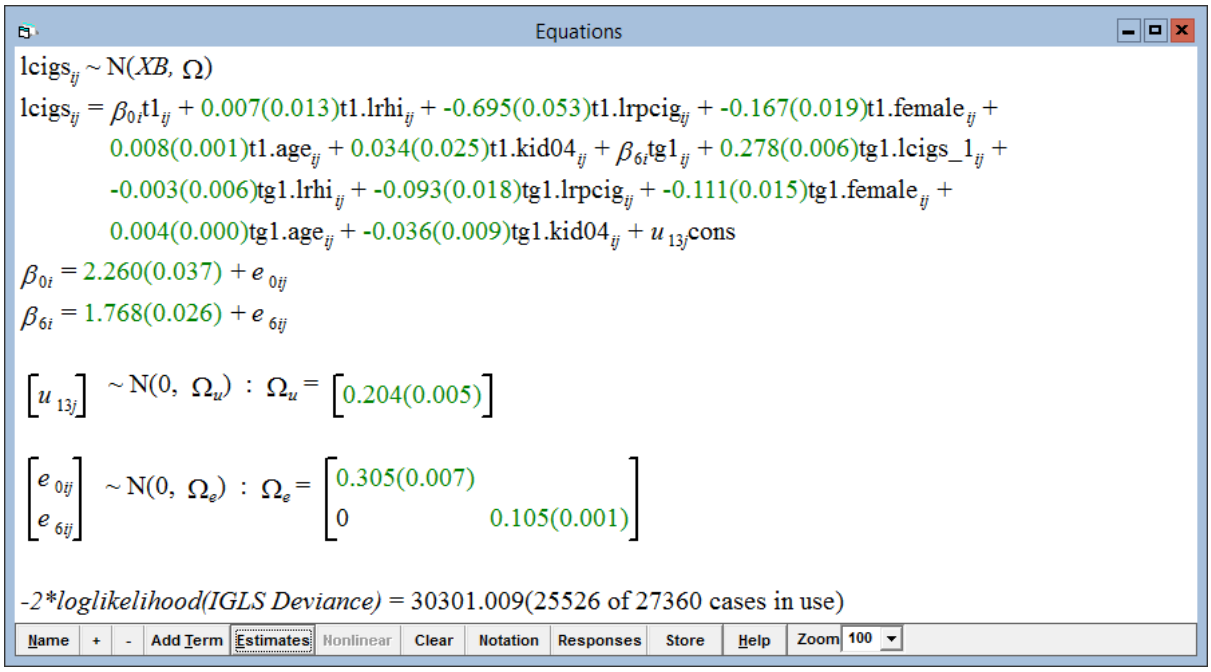
Name	Cn	n	missing	min	max	categorical	description
id	1	27360	0	1	5475	False	Person identifier
year	2	27360	0	1991	2002	False	Year of interview
ncigs	3	27360	0	1	81	True	usual no. of cigarettes smoke...
lcigs	4	27360	0	0	4.394449	False	ln(number cigarettes smoked)
lcigs_1	5	27360	1677	0	4.394449	False	ln(number cigs), lag(1)
lrpcig	6	27360	8	-8.175502E-03	0.5349713	False	ln(real price of cigarettes)
lrhi	7	27360	174	-1.267545	4.641064	False	ln(real household income)
age	8	27360	0	16	70	False	age at date of interview
female	9	27360	0	0	1	False	
kid04	10	27360	0	0	1	False	Has at least 1 child aged 0-4 y...
cons	11	27360	0	1	1	False	
c12	12	27360	0	1	12	False	
t1	13	27360	0	0	1	False	
tg1	14	27360	0	0	1	False	
c15	15	0	0	0	0	False	
c16	16	0	0	0	0	False	

The interactions between these new and modified variables are included as predictors in our model and the constant is excluded. A random intercept is specified at level 2, and separate residual variances for groups defined by **t1** and **tg1** (i.e. occasion 1 vs subsequent occasion) are fitted.

- From the **Model** menu, select **Equations**
- Click on the red **y** and select **lcigs** as the **y** variable, **2-ij** as the **N levels** variable, **id** as the **level 2(j)** variable, and **year** as the **level 1(i)** variable
- Click on **Add Term** and select **t1** in the variable drop-down box then click **Done**
- Click on **Add Term**, select **1** from the **order** drop-down box, **t1** in the first **variable** drop-down box, **lrhi** in the second **variable** drop-down box, and then click **Done** to add the interaction between **t1** and **lrhi**
- Continue to add interactions between **t1** and the following variables: **lrpcig**, **female**, **age**, and **kid04**
- Now add the variable **tg1** and interactions between **tg1** and the following variables: **lcigs_1**, **lrhi**, **lrpcig**, **female**, **age**, and **kid04**
- Add the variable **cons** to the model
- Click on the variable **t1**, check the **i(year)** check box and click **Done**. Now do the same for the **tg1** variable
- Click on the variable **cons**, uncheck the **Fixed parameter** check box, check the **j(id)** check box, and click **Done**
- Click the + button twice to show the full model specification
- Finally, click the term σ_{e06} in the covariance matrix and click **Yes** in the **remove term t1\tg1 from level 1 covariance matrix** pop up box
- Click **Estimates** to show the full model specification, check that your model appears as in the first screenshot below and then click **Start** to run the model

- Once the model has converged, click **Estimates** to reveal the parameter estimates and standard errors





For ease of comparison, selected estimates from Models 1 and 2 are given in the table below.

	Model 1		Model 2	
Parameter	**Est.**	**St. Err.**	**Est.**	**St. Err.**
Lagged response $y_{i-1,j}$ (γ)	0.473	0.006	0.278	0.006
Between-individual variance (σ^2_u)	0.085	0.003	0.204	0.005

Coefficient of individual random effect (λ)	-	-	1[a]	-

As expected, ignoring the initial condition leads to overstatement of state-dependence (γ) and understatement of unobserved heterogeneity (σ_u^2).

Model 3

An additional model (Model 3) is described in the Stata practical in which λ was unconstrained rather than constrained to the value of 1 as in Model 2. However, this model specification cannot be fitted in MLwiN and as such we do not cover it here. The results of Model 3 however suggest that constraining the value of λ to 1 as in Model 2 is not as efficient as allowing it to be freely estimated.

P15.9 Advanced Topics

There is no practical exercise for this section.

References

Madden, D. (2008) Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, 27(2): 300-307.

Steele, F., and Durrant, G. B. (2011) Alternative Approaches to Multilevel Modeling of Survey Noncontact and Refusal. *International Statistical Review*, **79**: 70-91.